Combining Psychological Theory with Language Models for Suicide Risk Detection

Daniel Izmaylov[1]
izmaylov@post.bgu.ac.il

Amir Bialer[1]
amirbial@post.bgu.ac.il

Avi Segal[1]
avisegal@gmail.com

Meytal Grimland[2]
meytal.grimland@gmail.com

Yossi Levi-Belz[2]
yossil@ruppin.ac.il

Kobi Gal[1,3]
kobig@bgu.ac.i

[1]Ben-Gurion University of the Negev [2]Ruppin Academic Center [3]University of Edinburgh

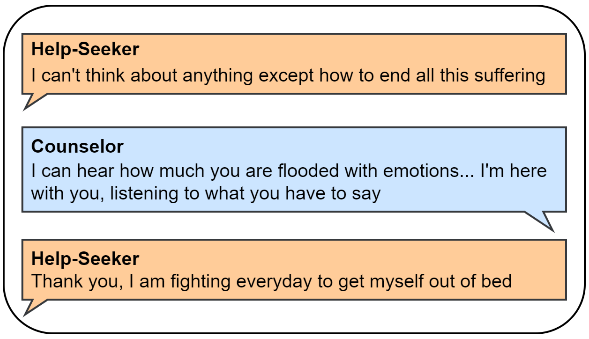
Abstract

Recent years saw a dramatic increase in the popularity of online counseling services providing emergency mental health support. This paper provides a new language model for automatic detection of suicide risk in online chat sessions between help-seekers and counselors. The model adapts a hierarchical BERT language model for this task. It extends the state of the art in capturing aspects of the conversation structure in the counseling session and in integrating psychological theory into the model. We test the performance of our approach in a leading national online counseling service that operates in the Hebrew language. Our model outperformed other non-hierarchical approaches from the literature, achieving a 0.76 F2 score and 0.92 ROC-AUC. Moreover, we demonstrate our model's superiority over strong baselines even early on in the conversation, which is key for real-time detection in the field. This is a first step towards incorporating suicide predictive models in online support services and advancing NLP tools for resource-bounded languages.

1 Introduction

Suicide accounts for more than 700,000 lives lost across the world every year. It is the second leading cause of death for adolescents and adults from 15 to 29 years of age in many countries. A key effort in suicide prevention is to identify individuals at risk of suicide as early as possible (World-Health-Organization, 2021).

In the past decade, online counseling services for mental health support have become commonplace in many countries, providing chat support and guidance to at-risk individuals (see fictitious example in Figure 1). Online counseling services aim to provide mental support and address a variety of mental health crises through specialist counselors. These counselors are trained to detect suicide risk during conversations and intervene quickly as needed. These services have experienced tremendous growth in traffic since the commencement of the COVID pandemic (Zalsman et al., 2021). Any kind of technological support to help counselors in this critical task can potentially save lives.



Help-Seeker
I can't think about anything except how to end all this suffering

Counselor
I can hear how much you are flooded with emotions... I'm here with you, listening to what you have to say

Help-Seeker
Thank you, I am fighting everyday to get myself out of bed

Figure 1: A fictitious example of a conversation.

This paper provides a computational model for detection of suicide risk from anonymous text-based discussions between help-seekers and counselors. Our data is taken from an online counseling service in a low-resource language (Hebrew). There are several challenges towards solving suicide risk detection in our setting: State of the art pre-trained language models for suicide prevention usually focus on posts from social media which are very different in structure from online conversations between counselors and help-seekers. Existing works that do consider conversations in this domain ignore the conversation structure or are limited in the size of the conversation they consider. Also, the set of NLP resources available for low-resource languages is extremely limited when compared to English.

To address this gap, we present a hierarchical language model called SR-BERT that includes a base layer for encoding the conversation text and an additional layer for capturing aspects of conversation structure. The hierarchical structure of SR-BERT encodes each of the messages in the conversation separately and is not limited by the size

Findings of the Association for Computational Linguistics: EACL 2023, pages 2430–2438
May 2-6, 2023 ©2023 Association for Computational Linguistics

of the conversation. We hypothesized that incorporating knowledge from suicide risk theory as part of the pre-training step can improve downstream performance of the detection model. To this end, we develop a new domain knowledge-based pre-training step that embeds a Suicide Risks Factor lexicon (SRF) into SR-BERT. The SRF lexicon was created by a team of psychologists who are experts on suicide risk theory and prevention.

In empirical studies, SR-BERT significantly outperforms alternative classifiers for suicide risk (SR) detection, including the state-of-the-art (Bialer et al., 2022). We show that adding the domain-expert information to SR-BERT plays a critical part in its performance. In particular, it obtained consistently better performance than Bialer et al. (2022) when processing different portions of the conversation. These findings suggest that SR-BERT can perform well in the field when analyzing conversations in real-time.

We extend the state-of-the-art hierarchical language models to combine conversation structure and expert-based knowledge in the pretraining step. We show this approach leads to significant increases in performance for detecting suicide risk from chat conversations.

2 Related Work

This paper relates to past studies in suicide risk detection in online settings, representing domain knowledge and conversation structure in deep language models and NLP tools for low-resource languages. We expand on each of these topics in turn. For a review on using machine learning in suicide prevention, we refer the reader to Ji et al. (2021).

The majority of work using machine learning to predict suicide risk analyzes posts from social media (Coppersmith et al., 2018; Zirikly et al., 2019; Shing et al., 2018; Sawhney et al., 2018; Tadesse et al., 2019). Recent works in this social media suicide prediction space includes Cao et al. (2019) who used an LSTM with an additional attention layer to predict SR from social media posts, and Wang et al. (2021) who combined a generic BERT model with predefined rules for scoring suicide risk in social media. Additionally, Ophir et al. (2020) showed that psychological questionnaires can improve the performance of neural networks to identify at-risk individuals from Facebook posts. We significantly differ from the social media setting in our focus on conversations from online counseling services, where messages are significantly longer than social media posts, and messages are part of a conversational structure and exhibit psychological dynamics. We show that capturing these aspects in the conversation model is necessary for recognizing SR in our setting.

There are few works on suicide detection in online counseling conversations, but none of these reasons about the conversation structure in the session. Most relevant to our approach is the model by Bialer et al. (2022) who combined a pre-trained language model based on BERT (Devlin et al., 2018) with a lexicon of suicide terms that were manually extracted from conversations. This model was able to represent only part of the conversation (512 tokens) and ignored the input from the counselor. Our SR-BERT model used a lexicon extracted from psychological theory, which was embedded in the pre-training process. The model significantly outperforms that of Bialer et al. (2022) on the same dataset, both for entire conversations as well as when considering early detection on parts of the conversation.

We mention two approaches for detecting SR in counseling services that did not consider early detection. Xu et al. (2021) combined a word2vec representation of suicide concepts with a bi-directional LSTM network for SR prediction in Korean online counseling service. Each side of the conversation was represented by an independent BI-LSTM. This approach used a knowledge graph to represent a psychological lexicon which may be more time-consuming for human experts to construct. Our model is shown to outperform a baseline using a similar representation (doc2vec) on our dataset. Bantilan et al. (2021) used TF-IDF embedding with XGBoost to predict SR in transcribed phone calls from an English counseling service. This model did not use a lexicon.

There is ample evidence on the benefits of incorporating domain knowledge in language models for downstream tasks (Childs and Washburn, 2019; Cao et al., 2019; Lee et al., 2020; Colon-Hernandez et al., 2021; Gaur et al., 2019). Notable examples include Gaur et al. (2019) and Wang et al. (2021) who showed that using lexicon-based features can improve machine learning prediction of suicide risk in Chinese blogs. They use lexicons to map terms from online discussions to clinically relevant sets of categories. We extend these approaches by presenting a new method for incorporating domain knowledge in the pre-training phase of deep learning models.

In general, NLP models and solutions for low-resource languages are extremely limited. In Hebrew, two pre-trained language models were published, HeBERT(Chriqui and Yahav, 2021) and AlephBERT (Seker et al., 2022). We used AlephBERT which is freely available and was trained on a larger dataset than HeBERT and was able to outperform HeBERT on a variety of natural language tasks. We are first to use hierarchical transformer architecture to model conversation structures in a low-resource language.

3 The Sahar domain

Sahar (Hebrew acronym for Online Mental Health Support [1]) was established in 2000 and is the leading internet-based emotional support and suicide prevention organization in Israel. It provides anonymous, confidential, and free crisis support via a chat hotline (in Hebrew and in Arabic). The organization handles more than $40,000$ chat sessions per year, and these numbers have increased significantly during the COVID-19 pandemic (Zalsman et al., 2021).

Sahar counselors are volunteers who receive year-round guidance and supervision from a team of mental health professionals. Shifts take place in the evening hours and are accompanied by trained therapists who monitor the conversations and provide professional support to counselors as needed. During the shifts, counselors work in a high-stress environment and usually handle multiple chat sessions in parallel at any given time. Counselors provide a written summary of each of their conversations, as well as indicate whether the conversation exhibits suicide risk.

The Sahar corpus contains more than $40,000$ chat sessions (conversations) that took place over the span of five years (2017-2022). Each conversation includes the messages generated by the help-seekers and the counselors, ordered by time signatures. Table 1 presents general statistics about the dataset. We note that 39.5% of the sessions are labeled with either positive or negative SR label and 17% of these sessions are SR positive.

Table 1: General statistics for Sahar corpus

Total num. of sessions	44,506
Num. of labeled sessions	17,564
SR positive label ratio	17%
Mean(Median) num. of messages	57(46)
Mean(Median) num. of turn exchanges	27(25)
Mean(Median) num. of tokens	617(566)

To validate the SR labels, a sample of 600 conversations (300 positive SR, 300 negative SR) was labeled separately by clinical psychologists with expertise in suicide theory. The Krippendorff's α for inter-annotator agreement between the psychologists and the SR label in the conversation is 0.766, which is en par with other works. We note that the inconsistencies found in the samples were debated by the psychologists and resolved in the data set.

4 The SRF Psychological Lexicon

As part of our research, a team of psychology experts from a national center for suicide prevention in Israel has constructed a Suicide-Risk Factors Lexicon (SRF) in Hebrew that is based on psychological theory.

The SRF lexicon contains terms relating to personal and situational variables associated with an increase in suicidal thinking, based on valid self-report questionnaires in the psychological and psychiatric literature (Klonsky and May, 2015; Turecki and Brent, 2016; Nock et al., 2008).

Each of the 3,094 sentences in the lexicon belonged to one of 25 categories. Specifically, terms relating to depression are taken from the Patient Health Questionnaire Depression Module (PHQ-9) (Kroenke et al., 2001). Terms relating to a sense of burdensomeness are taken from the Interpersonal Needs Questionnaire (INQ) (Van Orden et al., 2012). Terms relating to a sense of hopelessness are taken from the Beck hopelessness scale (Beck et al., 1996). Terms relating to suicide behavior were taken from the Columbia questionnaire (Posner et al., 2008) which is a standard tool to measure suicide risk.

Examples of sentences for the category "perceived burdensomeness" (translated) included sentences such as "better without me", "I am a burden", "I spoil everything for my spouse"; and the lexicon category "explicit suicide mentions" contains phrases such as: "to die", "to commit suicide", "kill myself" etc.

5 The SR-BERT Language Model

Our main contribution is SR-BERT, a two-layer hierarchical language model that extends the generic DialogBERT (Gu et al., 2021) to reason about con-

[1] https://sahar.org.il

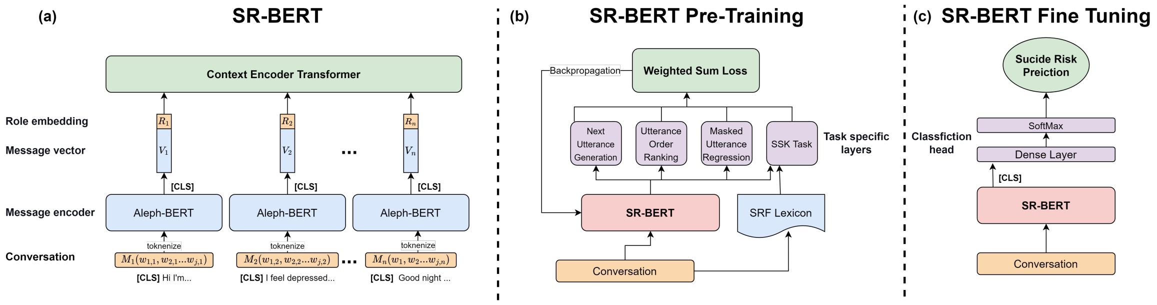


Figure 2: Model architecture. (a) SR-BERT base architecture, encoding conversation and speaker roles. (b) Pre-training procedure on 4 self-supervised tasks including psychological knowledge learning using the SRF lexicon. (c) Fine-tuning procedure learning to predict Suicide Risk (SR)

versation structure in suicide risk prediction settings and harness psychological domain knowledge. The SR-BERT architecture is shown in Figure 2(a).

The architecture is composed of two part: A transformer based layer performing message encoding, and on top of it an additional transformer layer, which captures conversation structure, named Context Encoder Transformer.

The base layer uses the AlephBERT (Seker et al., 2022) pre-trained language model to encode each message in the dialogue to a vector. The received message encoding is then combined with speaker role representation (help-seeker vs. counselor) to capture important conversation aspects such as turn-taking. The Context Encoder Transformer is a transformer based encoder applied at the message level (instead of the single token level) which transforms the series of message vectors into a context-sensitive repression of the conversation. The Context Encoder Transformer included 12 attention layers, and 12 hidden layers, each with a vector size of 780. The hidden layer size is 780 rather than 768 in AlephBert to account for the additional speaker role encoding.

The hierarchical structure of the architecture enables the model to capture multiple messages including turn exchanges and speaker roles. Furthermore, it enable the encoding of each message independently, thus avoiding the need to truncate conversations (due to AlephBert's 512 token limit) as in past work.

5.1 Pre-training with Self Supervised Knowledge

In this section we describe the use of several pre-training tasks for adapting SR-BERT to conversation structure of online counseling, including a new pre-training task for incorporating the SRF lexicon. This procedure uses the entire Sahar dataset, and is shown in Figure 2 (b).

The first step in this process is to represent conversations as a 25 dimension vector representing the different categories in the lexicon. For a given conversation, the value at index k is the number of sentences in the conversation with at least one occurrence in the kth lexicon category.

We also considered a reduced 5-dimension representation of conversations on the SRF lexicon space. To this end we selected the top categories using XGBoost feature selection (Chen and Guestrin, 2016) on the SR prediction task of entire conversations. We identified the top 5 categories as "self perceived burdensomeness", "previous suicide attempt", "loss of hope" "self injury" and "suicidal thinking". The 5-dimension representation outperformed the 25-dimension representation on the validation set, leading us to use this representation in the subsequent pre-training phase.

The second step, called the Self Supervised Knowledge task, applies a new pre-training task for predicting Sahar conversations in the SRF representation space. For a given prefix of a conversation, we mask a message in this subset with a fixed probability of 80%. We then use SR-BERT to predict the conversation subset's representation in the SRF space using a fully connected layer. The loss is obtained by calculating the mean squared error (MSE) between the original subset representation and the predicted (masked) representation in the SRF space. This process is repeated for increasing size of conversation prefixes, to simulate conversations of varying sizes.

In addition to the SSK task, we implemented the three pre-training tasks defined by DialogBERT(Gu et al., 2021) for capturing several aspects of the conversation structure: message-level semantics, conversation structure, and underlying dialogue sequential order. We describe them briefly here and refer the reader to the full paper for more details.

- **Next Utterance Generation** The goal of this task is to generate the next message in the conversation when the previous messages are given. The task tries to minimize the cross-entropy loss between the predicted words and the original words of the next message.

- **Masked Utterance Regression** The goal of this task is to predict a randomly masked message in a conversation from its context. The loss is obtained by calculating the MSE between the original and the predicted message vectors.

- **Distributed Order Ranking Network** This task predicts the order index of each message from a shuffled order of a conversation. The task tries to minimize the KL divergence between the predicted order and the true order.

The calculated loss for the model propagation over the four self supervised tasks is the weighted sum of each loss function in the pre-training stage. The AdamW optimizer is employed with a linear planned warm-up technique and an initial learning rate of 5e-5. Additionally, we use an adaptive learning-rate scheduler with 0.01 weight decay, 15,000 warm-up steps, and a batch size of 32. The model is trained for 20 epochs. All experiments are conducted on a GeForce RTX 3090 GPU using the PyTorch package.

5.2 Fine-tuning

In the fine-tuning step Figure 2(c), SR-BERT is adapted for the suicide risk prediction task using a standard approach (Sun et al., 2020). To this end we add a binary classification head to SR-BERT. The classification head consists of a dense layer with an output size of 2 and a softmax activation function. By maximizing the log-likelihood of the actual label, we fine-tune the Context Encoder Transformer and the classification head. We employ the AdamW optimizer with a linear planned warm-up technique and an initial learning rate of 2e-5. Additionally, we use an adaptive learning-rate scheduler with 0.01 weight decay, and a batch size of 16. The model is trained for 10 epochs.

6 Empirical Methodology

We randomly split the labeled Sahar dataset to a train (70%) validation (15%) and test (15%) sets. These data sets were used throughout the experiments described in the following section. The validation set was used for training model hyper parameters.

We follow prior work in evaluating model performance using ROC-AUC which is widely employed in suicide detection research (Bernert et al., 2020). Additionally, we report on the F2-score (Sokolova et al., 2006) for predicting the positive SR label. This measure concentrates on reducing false negatives (rather than false positives) and is thus well suited for SR detection where missing a positive class has life threatening implications.

We compare SR-BERT with SSK to the following baseline models:

6.1 SR-BERT w.o.SSK

This model omits the SSK pre-training task from SR-BERT w. SSK. Apart from the SSK pre-training task this model is identical to SR-BERT w. SSK. including the hierarchical structure and pre-training on the other 3 tasks.

6.2 Explicit based lexicon + XGBoost

We used an XGBoost classifier that was based on an encoding of conversations over the explicit suicide related terms proposed by Bialer et al. (2022). This list includes 67 terms such as “commit suicide”, “cut wrists”, “wish to die” etc. We note that explicit terms carry very weak signal for SR detection.

6.3 Ensemble SI-BERT (Bialer et al., 2022)

This is a non-hierarchical Hebrew language model ensembled with a classifier based on the Explicit lexcion, that represents the state of the art for SR detection. It was trained on the same dataset from the Sahar organization. To bypass BERT’s constraint of 512 tokens, Ensemble SI-BERT only utilized the help seeker text and truncated text greater than 512 tokens. We re-implemented this model with the code and parameters provided by the authors and run it on the dataset provided for this research. This is the reported state of the art for this domain in the Hebrew language.

6.4 SRF based lexicon + XGBoost

An XGBoost (Chen and Guestrin, 2016) classifier based on the 5-dimension SRF conversations rep-

Table 2: SR prediction results of compared models. Bold highlights highest value.

Model	Recall [%]	Precision [%]	ROC-AUC [%]	F2 [%]	F1 [%]
Doc2Vec+XGBoost	31.3	69.2	64.7	35.1	43.1
Explicit lexicon+XGBoost	49.2	67.1	76.9	52.3	57.7
SRF lexicon + XGBoost	55.1	67.2	76.5	57.1	60.0
Ensemble SI-BERT	60.4	**70.9**	91.3	62.3	65.3
SR-BERT w.o. SSK	72.9	68.4	**92.1**	71.9	70.6
SR-BERT w. SSK	**78.3**	68.9	**92.1**	**76.2**	**73.3**

resentation over the SRF lexicon. We note that XGBoost outperformed Random Forest and Logistic Regression as the classifier for this baseline (and for the next two baselines)

Consider for example one of the sessions which includes the statement “I am having strong stomach aches since yesterday, I want to die.”. This session includes a term from the Explicit lexicon while it is not an SR positive session.

6.5 Doc2Vec + XGBoost

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

7 Results

We first present the performance of the SR-BERT model in predicting SR on labeled conversations compared to the proposed baselines. Results are then reported for early SR detection, when increasing percentages of conversation information are available.

7.1 SR Detection from Complete Conversation

Table 2 compares the performance of the SR-BERT model to the baselines when predicting suicide risk from complete conversations. As seen in the table, both SR-BERT-based models (with and without SSK pre-training) outperformed the Ensemble SI-BERT model in terms of recall, F1, F2, and ROC-AUC metrics. Most notable improvement was in the recall metric where SR-BERT w.o. SSK achieved a 12.5% improvement over the Ensemble SI-BERT model, which led to a 9.6% improvement in the F2 metric. Moreover, the additional SSK pre-training improved on the SR-BERT w.o. SSK results for all metrics except the ROC-AUC score, where it hasn’t change. Ensemble SI-BERT achieved the highest precision, which was slightly better than SR-BERT w. SSK. It exhibited a substantially lower recall score, which correlates to lower F1 and F2 values.

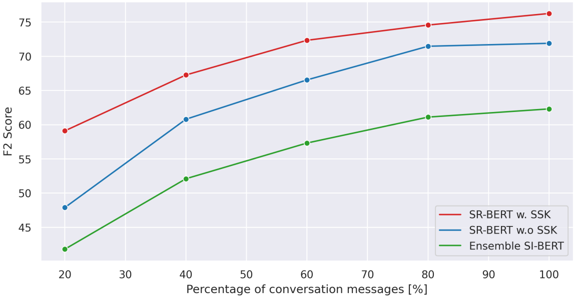


Figure 3: Classification results for early detection of top-performing SR detection approaches

The SRF lexicon + XGBoost based classifier was better than the Explicit lexicon + XGBoost classifier in all measures apart from ROC-AUC. We also note that the BERT based models outperformed the none BERT models on all tested metrics.

We used the McNemar paired test for labeling disagreements (Gillick and Cox, 1989) to compare between the predictions of the different models. Statistical significance with $p < 0.05$ was demonstrated for SR-BERT w. SSK vs. SR-BERT w.o. SSK and for SR-BERT w. SSK vs. Ensemble SI-BERT.

Overall SR-BERT w. SSK achieved a substantial improvement in recall and F2 compared the Ensemble SI-BRET of 17.9% and 13.9% respectively, with only a slight decrease in precision performance. This is critical in the suicide risk detection realm where recall is key to identifying help-seekers at risk and enabling targeted support.

7.2 Early SR Detection

Evaluating the ability of SR-BERT to predict SR risk from partial sessions provides an indication of its performance in real time, when only part of the session is available. To this end, Figure 3 compares the performance of the different models after receiving the first $\{20, 40, 60, 80, 100\}$ percent of messages in the session. As seen in the figure, the performance of all models improved as the sessions progressed. However, SR-BERT w. SSK model consistently outperformed the other models, followed by SR-BERT w.o. SSK. The difference in performance between SR-BERT with SSK and SR-BERT w.o. SSK was the largest at the beginning of the session and reduced as the sessions advanced. This may indicate the contribution of SR-BERT w. SSK to identify risk variables from the lexicon in early stages of the dialogue when information is lacking. In contrast, the difference in performance between SR-BERT w.o. SSK and the Ensemble SI-BERT model increases as sessions advance. This could be due to the inability of Ensemble SI-BERT to process the lengthy dialogue without having to truncate it, which may result in the loss of important information as sessions develop.

8 Conclusion and Future Work

This work has provided a new automatic approach for suicide risk detection in online conversations between help-seekers and counselors. Early detection of at-risk individuals is a key goal of suicide prevention. Our approach extends the state-of-the-art in deep language modeling by 1) incorporating domain knowledge relevant to suicide risk detection as part of the pre-training step; 2) reasoning about the structure of the conversation between help-seekers and counselors; 3) adapting to a low-resource language (Hebrew). The presented approach was able to significantly outperform the state-of-the-art approaches when detecting SR from complete conversations, as well as early detection when only part of the conversation is available. These results suggest the model may be able to support the work of counselors in real chat sessions, alerting them in real-time to at-risk individuals and enabling quick and focused responses. For future work, we intend to improve our approach by capturing more aspects of conversations, such as prosody (Wilson and Wharton, 2006; Kliper et al., 2010) as well as model the mental state dynamics of the help-seeker. We are also extending the model with explanations to be able to provide justifications for predictions made and point to key exchanges and phrases that triggered specific predictions.

9 Limitations

We note several limitations of this study.

First, our model was evaluated only in the Hebrew language. We have not directly compared SR-BERT to approaches for detecting suicidal risk in non-Hebrew domains, and note that the effectiveness of the model may vary across different languages and cultural contexts. It is difficult to make this comparison given the lack of public data sets from online counseling services.

Second, the proposed approach relies on the existence of psychological knowledge for pre-training the SR-BERT model which requires human effort. On the one hand, psychological lexicons already exist in English (Lee et al., 2020) and possibly in other languages. On the other hand, lexicons inherently suffer from limited coverage, lack of context and are expensive to maintain. Sharing domain knowledge across research tasks may go a long way to overcome these issues. We intend to make the lexicon developed for this research publicly available.

Third, the annotation of the help seekers' mental state was performed by the counselors, rather than the help seekers themselves. While the counselors underwent a thorough training process lasting several months and were monitored by certified clinical psychologists, there is still the possibility that they may have misclassified the mental state of the help seekers. This issue is prevalent in many studies that rely on observer-reported data.

Finally, the current model does not provide any explanations for its predictions, which are of high importance in order to support counselors in the field. This is essential in order to ensure that the model is not merely a means of classification but instead is able to provide valuable insights and assistance to counselors. This is a key focus of our future development plans.

10 Ethics Statement

The present study has been conducted in accordance with the highest ethical standards and has been approved by the relevant institutional review board of the participating institutions. All data utilized in this study, including the Sahar corpus of conversations between help-seekers and counselors, and the SRF psychological lexicon, have been obtained in compliance with the IRB. Specifically, the Sahar dataset has been anonymized and encrypted to protect the privacy of the participants, and all help-seekers who have provided data for this study have given informed consent for the anonymous use of their sessions for research purposes. The counselors signed consent papers to allow the usage of their text data for the study.

It is important to note that despite the model's ability to successfully predict SR during the conversation and its demonstrated gender fairness, it is not intended to replace human volunteer counselors. We believe that human involvement is essential in providing support to help-seekers, and the role of an automated model is to serve as an aid to counselors, enhancing their ability to assess SR rather than replacing them. Our take is that in the future, when such models could be deployed in the field (after all necessary approvals and adaptations), they may only act as a "friendly parrot" on the counselors' shoulders, providing additional insights and supporting their decision-making process in the high load situations these counselors are facing on a daily basis.

References

Niels Bantilan, Matteo Malgaroli, Bonnie Ray, and Thomas D. Hull. 2021. Just in time crisis response: Suicide alert system for telemedicine psychotherapy settings. 31(3):302–312.

Aaron T Beck, Robert A Steer, and Gregory Brown. 1996. Beck depression inventory–ii. *Psychological assessment*.

Rebecca A Bernert, Amanda M Hilberg, Ruth Melia, Jane Paik Kim, Nigam H Shah, and Freddy Abnousi. 2020. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *International journal of environmental research and public health*, 17(16):5929.

Amir Bialer, Daniel Izmaylov, Avi Segal, Oren Tsur, Yossi Levi-Belz, and Kobi Gal. 2022. Detecting Suicide Risk in Online Counseling Services: A Study in a Low-Resource Language. The 29th International Conference on Computational Linguistics (COLING-22) http://shorturl.at/crXY2.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Christopher M. Childs and Newell R. Washburn. 2019. Embedding domain knowledge for machine learning of complex material systems. 9(3):806–820.

Avihay Chriqui and Inbal Yahav. 2021. HeBERT & HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *The World Wide Web Conference on - WWW '19*, pages 514–525. ACM Press.

Laurence Gillick and Stephen J Cox. 1989. Some statistical issues in the comparison of speech recognition algorithms. pages 532–535.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12911–12919.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. 8(1):214–226.

Roi Kliper, Yonatan Vaizman, Daphna Weinshall, and Shirley Portuguese. 2010. Evidence for depression and schizophrenia in speech prosody. In *Proceedings of the 3rd ICSA Tutorial and Research Workshop on Experimental Linguistics 2010*, pages 85–88.

E David Klonsky and Alexis M May. 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.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. 32(2):1188–1196.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. pages 2208–2217.

Matthew K Nock, Guilherme Borges, Evelyn J Bromet, Christine B Cha, Ronald C Kessler, and Sing Lee. 2008. Suicide and suicidal behavior. *Epidemiologic reviews*, 30(1):133–154.

Yaakov Ophir, Refael Tikochinski, Christa S. C. Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. 10(1):16685.

K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. 2008. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Greenfeld, and Reut Tsarfaty. 2022. Alephbert: Language model pre-training and evaluation from sub-word to sentence level. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46–56.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the fifth workshop on computational linguistics and clinical psychology: from keyboard to clinic*, pages 25–36.

Marina Sokolova, Nathalie Japkowicz, Stan Szpakowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to Fine-Tune BERT for Text Classification?

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Gustavo Turecki and David A Brent. 2016. Suicide and suicidal behaviour. *The Lancet*, 387(10024):1227–1239.

Kimberly A Van Orden, Kelly C Cukrowicz, Tracy K Witte, and Thomas E Joiner Jr. 2012. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the interpersonal needs questionnaire. *Psychological assessment*, 24(1):197.

Rui Wang, Bing Xiang Yang, Yujun Ma, Peilin Wang, Qiao Yu, Xiaofen Zong, Zhen Huang, Simeng Ma, Long Hu, Kai Hwang, and Zhongchun Liu. 2021. Medical-Level Suicide Risk Analysis: A Novel Standard and Evaluation Model. 8(23):16825–16834.

Deirdre Wilson and Tim Wharton. 2006. Relevance and prosody. *Journal of pragmatics*, 38(10):1559–1579.

World-Health-Organization. 2021. Live life: an implementation guide for suicide prevention in countries.

Zhongzhi Xu, Yucan Xu, Florence Cheung, Mabel Cheng, Daniel Lung, Yik Wa Law, Byron Chiang, Qingpeng Zhang, and Paul S.F. Yip. 2021. Detecting suicide risk using knowledge-aware natural language processing and counseling service data. 283:114176.

Gil Zalsman, Yael Levy, Eliane Sommerfeld, Avi Segal, Dana Assa, Loona Ben-Dayan, Avi Valevski, and J John Mann. 2021. Suicide-related calls to a national crisis chat hotline service during the covid-19 pandemic and lockdown. *Journal of psychiatric research*.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. Clpsych 2019 shared task: Predicting the degree of suicide risk in reddit posts. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 24–33.