Combining Psychological Theory with Language Models for Suicide Risk Detection

Anonymous submission

Abstract

Recent years saw a dramatic increase in the popularity of online counseling services providing emergency mental 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 which 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 predictive models in online support services, potentially helping human caregivers provide better support for help-seekers.

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 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 suicide prevention have become commonplace in many countries, providing chat support and guidance to at-risk individuals. The premise of these online services is that specialist counselors can detect suicide risk during the conversation and intervene as quickly as possible. 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.

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 the suicide risk detection in our setting: State of the art pre-trained language models for suicide prevention are limited in the size of the conversation they consider, and they do not relate to the conversation structure. Also, the set of NLP resources available for low-resource languages is extremely limited when compared to English. Finally, existing language models do not embed expert knowledge on suicide risk identification, thus potentially missing important signals for detecting such risk.

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 of the conversation. We hypothesized that incorporating domain knowledge relevant to suicide risk detection as part of the pre-training step can guide model learning and improve downstream performance. 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 which 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). Adding the domain-expert information to SR-BERT plays a critical part in its performance. Moreover, it obtained consistently better performance than that of the state-of-the-art when processing different portions of the conversation. Our findings suggests that SR-BERT can achieve good performance in the field, when analyzing conversations in real time.

The contributions of our work are twofold. First, we extend the state of the art hierarchical language models to combine conversation structure and expert-based knowledge. We show this approach leads to significant increases in performance for detecting suicide risk from chat conversations. Second, we extend the set of NLP tools available for resource-bounded languages and are

making our code and the SRF lexicon available to the research community at large.

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 developing 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).

SR detection in online counceling and social media A handful of studies study suicide detection in online support sessions. None of them reason about the conversation structure in the session. Specifically, Xu et al. (2021) combined a word2vec representation of suicide concepts with an bi-directional LSTM network for SR prediction in social networks. Each side of the conversation was represented by an independent BI-LSTM. Bantilan et al. (2021) used TF-IDF embedding with XGBoost in transcribed phone calls from an English counseling service. Neither of the above approaches considered early detection of suicide risk.

Bialer et al. (2022) combined a pre-trained language model based on BERT (Devlin et al. 2018) with a manually created explicit suicide mentions lexicon to predict suicide risk in online counseling sessions. Their model was only able to represent 512 tokens of the conversation and ignored the input from the counselor. Such an approach might cause a model to misinterpret the text and overlook important exchanges between help-seeker and counselor, which are crucial in dialogue-based communication. We show in this paper that our SR-BERT model significantly outperforms Bialer et al. (2022) approach on the same dataset, for entire conversations as well as when considering early detection.

The majority of work using machine learning to predict suicide risk analyzes posts from social media. This is a significantly different setting than online counseling in that messages are short, rarely relate to other messages and lack the temporal and psychological dynamics that characterize discussions with counselors. Cao et al. (2019) and Lee et al. (2020) combined LSTM with an embedding of suicide factors lexicon to predict SR in social media in Chinese and Korean, respectively. Wang et al. (2021) used an ensemble of predefined rules for scoring suicide risk and a generic BERT model in Chinese. Ophir et al. (2020) used neural networks to identify at-risk individuals from Facebook posts and psychological tests.

Representing domain knowledge and dialogue in language models There is ample evidence on the benefits of incorporating domain knowledge in language models for downstream tasks (Childs and Washburn 2019; Colon-Hernandez et al. 2021). Gaur et al. (2019) and Wang et al. (2021) 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.

Recent studies have demonstrated substantial developments in conversation structure modeling. Examples of systems that modeled discourse-level exchanges include DialoGPT (Zhang et al. 2019), GODEL (Peng et al. 2022), and DialogBERT (Gu, Yoo, and Ha 2021). DialogBERT is a hierarchical transformer language model with state of the art performance in a wide range of discourse-related applications. We chose to adapt and apply this model to the Sahar setting due to its flexibility in replacing its pre-trained language model, which is especially useful in supporting low-resource languages. We extended DialogBERT by developing new domain-specific tasks and demonstrating the architecture's performance in a classification task.

NLP tools for low-resource languages 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). The AlephBERT model was trained on a larger dataset than HeBERT and was able to outperform HeBERT on a variety of natural language tasks. Thus AlephBERT was selected for our work. To the best of our knowledge this research is the first effort to use a hierarchical transformer architecture to model conversation structures in a low-resource language.

The Sahar Corpus

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 $10,000$ 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 by 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 as needed. Each shift lasts for 3 hours, and includes 4 to 5 counselors working in parallel, covering together 50 to 80 sessions. The average session length is 32 minutes based on past Sahar history. 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) which took place in the span of five years (2017-2022). Each conversation includes the

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

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)

messages generated by the help-seekers and the counselors ordered by time signatures. Table 1 presents general statistics about the dataset. Sahar sessions include 57 messages on average and 27 turn-changes between counselor and help-seeker. 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. Additionally, the mean and median number of tokens in Sahar conversations exceed AlephBERT's limit of 512 tokens (Seker et al. 2022).

The SRF Psychological Lexicon

As part of our research, a team of psychology experts from a national center for suicide prevention have constructed a Suicide-Risk Factors based Lexicon (SRF). The SRF lexicon contains terms relating to personal and situational variables that are 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). Specifically, terms relating to depression in the lexicon are taken from the Patient Health Questionnaire Depression Module (PHQ-9) (Kroenke, Spitzer, and Williams 2001). Terms relating to sense of burdensomeness are taken from the Interpersonal Needs Questionnaire (INQ) (Van Orden et al. 2012). Terms relating to sense of hopelessness are taken from the Beck hopelessness scale (Beck, Steer, and Brown 1996).

In order to examine the level of suicidal thinking, the psychology team used the Columbia questionnaire (Posner et al. 2008), a leading tool for measuring suicidal behavior and suicide ideation. All the questionnaires and derived lexicon entries were tested and verified by a team of psychologists who are experts in the topic of suicide. In total, the expert team used 25 key variables known to be related to suicidal thinking leading to 25 categories.

Overall the developed lexicon contained 1753 sentences in these 25 categories. For example the lexicon category "perceived burdensomeness" (translated) included sentences such as "better without me", "I am a burden", "I spoil everything to my spouse"; and the lexicon category "explicit suicide mentions" contains phrases such as: "to die", "to commit suicide", "kill myself" etc.

The SR-BERT Language Model

Our main contribution is SR-BERT, a two-layer hierarchical language model that extends the generic DialogBERT (Gu, Yoo, and Ha 2021) to reason about conversation structure in suicide risk prediction settings and harness psychological domain knowledge. The SR-BERT architecture is shown in Figure 1(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.

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 1 (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

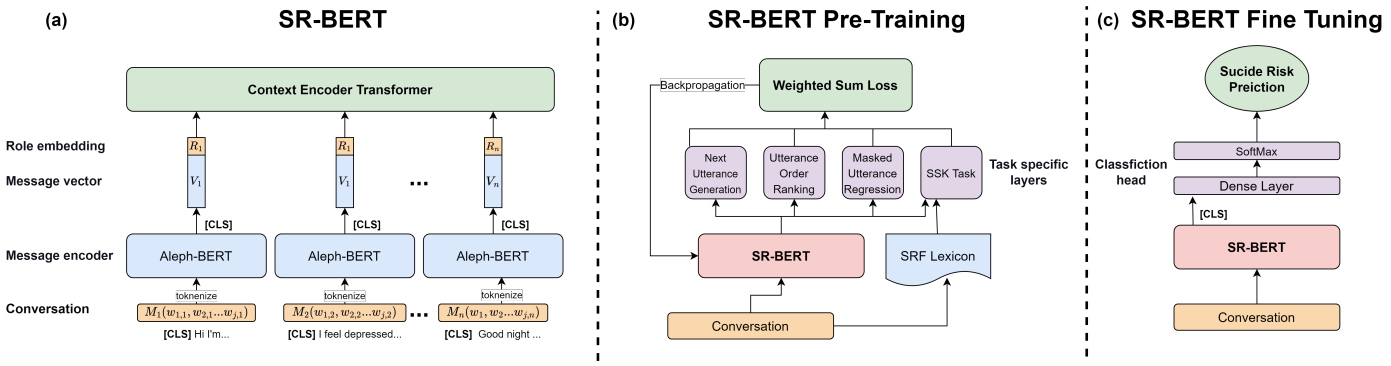

Figure 1: Model architecture. (a) SR-BERT base architecture, encoding conversation and speaker roles. (b) Pre-training procedure on 4 self-supervised tasks including psychological knowledge learning using the SRF lexicon. (c) Fine-tuning procedure learning to predict Suicide Risk (SR)

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, Yoo, and Ha 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 the PyTorch package.

Fine-tuning

In the fine-tuning step Figure 1(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.

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:

SR-BERT w.o. SSK This model omits the SSK pre-training task from SR-BERT w. SSK. Apart from the

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**

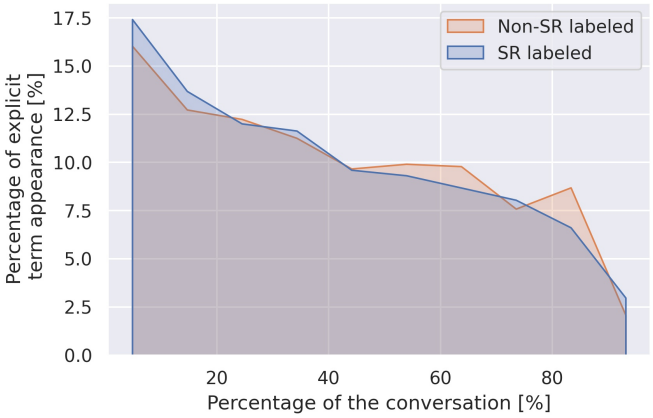

Figure 2: Distribution of explicit appearance of suicidal terms in the labeled Sahar dataset. Dark colors represent overlapping values.

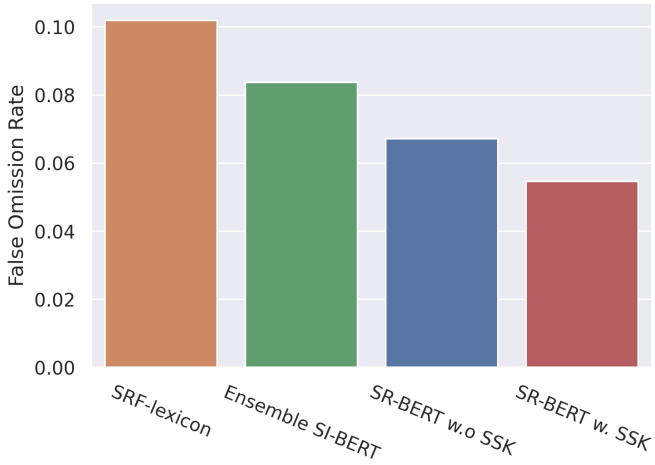

Figure 3: False Omission Rate of different models.

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.

Ensemble SI-BERT (Bialer et al. 2022) This is a non-hierarchical Hebrew language model 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.

SRF based lexicon + XGBoost An XGBoost (Chen and Guestrin 2016) classifier based on an encoding of conversations over the Suicide Risk Factor (SRF) lexicon. We note that XGBoost outperformed Random Forest and Logistic Regression as the classifier for this baseline (and for the next two baselines)

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. This is apparent in Figure 2 which presents the distributions of explicit terms throughout the session for SR and non-SR labeled conversations. As seen in the figure, the distributions over both classes are similar. 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.

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) .

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. Finally we look at the contribution of the SSK pre-training task for SR detection in the case of conversations where no explicit suicide related terms are present.

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.

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 SR-BERT w. SSK and the two models SR-BERT w.o SSK and Ensemble SI-BERT. 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 enable targeted support.

To further analyze this issue, Figure 3 compares the False Omission Rate for the different models, which is the fraction of false negative instances out of the set of all predicted negative instances. As seen in the figure, the SR-BERT models without SSK was able to reduce the False Omission Rate relative to the Ensemble SI-BERT Model from 0.083 to 0.067, a 19.2% improvement. Pre-training SR-BERT with the additional SSK task further reduced the False Omission Rate from 0.067 to 0.054, a total of 34% reduction compared to Ensemble SI-BERT.

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 4 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.

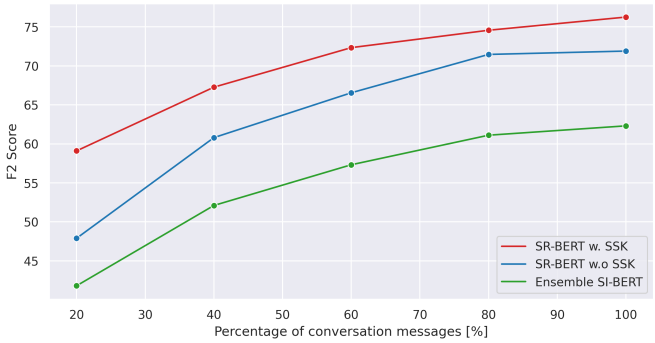

Figure 4: Classification results for early detection of top-performing SR detection approaches

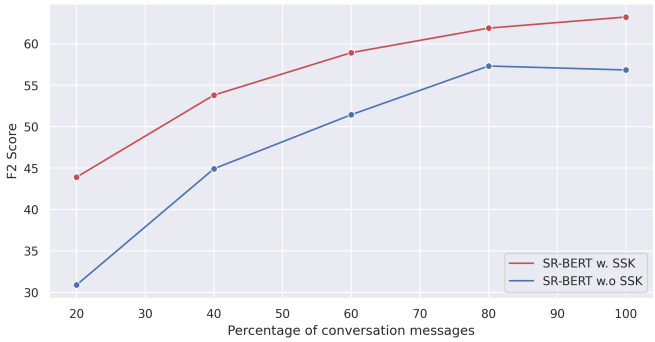

Figure 5: Classification results for early detection on explicit-less-terms benchmark.

To further assess the contribution of the SSK task, we compared the performance of the SR-BERT w. and w.o SSK on benchmark datasets that do not contain explicit mentions of suicide terms (from the Explicit based lexicon). Specifically, the "explicit-less-terms" benchmark dataset omits all explicit mentions of suicide terms from the testset, retaining the same number of messages. The "explicit-less-conversations" benchmark dataset omits all conversations with any mention of a suicide term from the Explicit lexicon.

The performance of SR-BERT with and without SSK

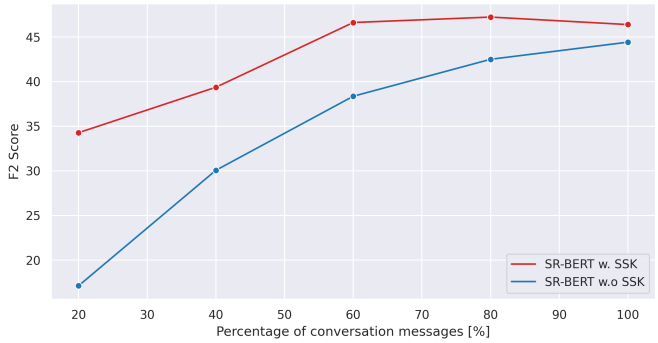

Figure 6: Classification results for early detection on explicit-less-conversations benchmark.

pre-training on the two benchmark datasets is presented in Figure 5 and Figure 6. As expected, the overall performance of both models decrease significantly compared to the original test-set, most notably on the explicit-less-conversations test. The model trained with the SSK task performed better on both benchmark datasets than the model trained without the SSK task for all session portions, with the major difference in performance apparent in the early stages of the support session. These results demonstrate the positive contribution of domain knowledge to suicide risk detection even for conversations which may not include explicit mentions of suicide terms.

Discussion

Our results demonstrate the importance of combining domain knowledge and conversation structure with pre-trained language models for the purpose of SR detection. In particular, embedding the SRF lexicon into SR-BERT captures nuances of the conversations that go beyond explicit mentions of suicide realted terms. Furthermore, the hierarchical SR-BERT approach was able to overcome the problem of limited input size that hinders existing approaches to language modeling. It was able to identify suicide risk variables early in the discussion and to consistently improve in performance as the conversations advanced, outperforming strong baselines. These results indicate that the model may be well suited for predicting suicide risk in real time, during actual support sessions.

We demonstrate the SR model's ability to capture nuanced textual exchanges with the two illustrative examples presented in Table 3. In the first example SR-BERT correctly labels the session as SR positive, while the Ensemble SI-BERT classifies it as negative SR (False Negative). Note that the suicide risk in this example becomes apparent after the clarification question by the counselor and by considering the response of the help-seeker in the context of their first message. This is exactly the contribution provided by the hierarchical approach of SR-BERT.

In the second example, SR-BERT correctly identified a non-SR session, while Ensemble SI-BERT mistakenly classified this session as SR (False Positive). We hypothesize that the flat ensemble model was misled by the "gun", "police" and "trauma" utterances while SR-BERT was able to weight these terms in context.

Finally, we reflect upon some limitations of this study. First, our model was evaluated only on the Hebrew language. We have not directly compared SR-BERT to approaches for SR detection in non-Hebrew domains. We note that public data sets from online counselling services are not available due to the inherent sensitivity of the materials.

Second, the approach relies on the existence of domain knowledge for pre-training the SR-BERT model that requires human effort. We note that psychological lexicons already exist in English (Lee et al. 2020). As we have shown, combining domain knowledge provides a remarkable enhancement for SR prediction. Sharing domain knowledge across research tasks may go a long way to facilitate future research in this area. We intend to make the lexicon developed for this research publicly available.

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 response.

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. Finally, in cooperation with the Sahar organization, we are planning to deploy SR-BERT as a support tool for counselors in the field, alerting them in real time to at-risk individuals.

Ethics Statement

All the data used in this study, including the Sahar corpus of conversations between help-seekers and counselors, and the SRF psychological lexicon, comply with

Table 3: Illustrative exchanges from conversations where SR-BERT and Ensemble SI-BERT disagree.

Help Seeker:	“I generally feel it’s over for me. ”
Counselor	“I’m not sure what you mean here dear, do you feel you wish to die? ”
Help Seeker:	“There is only one thing stopping me at this stage from doing this. ”
Counselor:	“If I understood correctly, you slapped your sister and locked yourself in the toilet. ”
Help Seeker:	“Correct. ”
Counselor:	“And you are afraid that the follow-up police visit will prevent you from shooting your gun at the shooting range? ”
Help Seeker:	“I am less worried about using my gun. I suffer and in trauma from the police visit itself. ”

IRB criteria in the relevant institutions. In particular, the Sahar data set is anonymized and encrypted, and all help-seekers providing data to this study have agreed to allow their sessions to be used anonymously for research purposes. This is part of the registration process in the Sahar website.

We consider a multi-level jury problem in which experts We consider a multi-level jury problem in which

experts We consider a multi-level jury problem in which
experts We consider a multi-level jury problem in which
experts

References

Bantilan, N.; Malgaroli, M.; Ray, B.; and Hull, T. D. 2021. Just in Time Crisis Response: Suicide Alert System for Telemedicine Psychotherapy Settings. 31(3): 302–312.

Beck, A. T.; Steer, R. A.; and Brown, G. 1996. Beck depression inventory–II. Psychological assessment.

Bernert, R. A.; Hilberg, A. M.; Melia, R.; Kim, J. P.; Shah, N. H.; and 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.; and Gal, K. 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.

Cao, L.; Zhang, H.; Feng, L.; Wei, Z.; Wang, X.; Li, N.; and He, X. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention.

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

Childs, C. M.; and Washburn, N. R. 2019. Embedding Domain Knowledge for Machine Learning of Complex Material Systems. 9(3): 806–820.

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

Colon-Hernandez, P.; Havasi, C.; Alonso, J.; Huggins, M.; and Breazeal, C. 2021. Combining Pre-Trained Language Models and Structured Knowledge.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gaur, M.; Alambo, A.; Sain, J. P.; Kursuncu, U.; Thirunarayan, K.; Kavuluru, R.; Sheth, A.; Welton, R.; and Pathak, J. 2019. Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention. In The World Wide Web Conference on - WWW '19, 514–525. ACM Press. ISBN 978-1-4503-6674-8.

Gillick, L.; and Cox, S. J. 1989. Some statistical issues in the comparison of speech recognition algorithms. 532–535.

Gu, X.; Yoo, K. M.; and Ha, J.-W. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 12911–12919.

Ji, S.; Pan, S.; Li, X.; Cambria, E.; Long, G.; and Huang, Z. 2021. Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. 8(1): 214–226.

Kliper, R.; Vaizman, Y.; Weinshall, D.; and Portuguese, S. 2010. Evidence for depression and schizophrenia in speech prosody. In Proceedings of the 3rd ICSA Tutorial and Research Workshop on Experimental Linguistics 2010, 85–88.

Klonsky, E. D.; and 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.

Kroenke, K.; Spitzer, R. L.; and Williams, J. B. 2001. The PHQ-9: validity of a brief depression severity measure. Journal of general internal medicine, 16(9): 606–613.

Le, Q.; and Mikolov, T. 2014. Distributed Representations of Sentences and Documents. 32(2): 1188–1196.

Lee, D.; Park, S.; Kang, J.; Choi, D.; and Han, J. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. 2208–2217.

Nock, M. K.; Borges, G.; Bromet, E. J.; Cha, C. B.; Kessler, R. C.; and Lee, S. 2008. Suicide and suicidal behavior. Epidemiologic reviews, 30(1): 133–154.

Ophir, Y.; Tikochinski, R.; Asterhan, C. S. C.; Sisso, I.; and Reichart, R. 2020. Deep Neural Networks Detect Suicide Risk from Textual Facebook Posts. 10(1): 16685.

Peng, B.; Galley, M.; He, P.; Brockett, C.; Liden, L.; Nouri, E.; Yu, Z.; Dolan, B.; and Gao, J. 2022. GODEL: Large-Scale Pre-Training for Goal-Directed Dialog. arXiv preprint arXiv:2206.11309.

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

Seker, A.; Bandel, E.; Bareket, D.; Brusilovsky, I.; Greenfeld, R.; and Tsarfaty, R. 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), 46–56.

Sokolova, M.; Japkowicz, N.; Szpakowicz, S.; and Szpakowicz, S. 2006. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence, 1015–1021. Springer.

Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2020. How to Fine-Tune BERT for Text Classification?

Turecki, G.; and Brent, D. A. 2016. Suicide and suicidal behaviour. The Lancet, 387(10024): 1227–1239.

Van Orden, K. A.; Cukrowicz, K. C.; Witte, T. K.; and Joiner Jr, T. E. 2012. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the Interpersonal Needs Questionnaire. Psychological assessment, 24(1): 197.

Wang, R.; Yang, B. X.; Ma, Y.; Wang, P.; Yu, Q.; Zong, X.; Huang, Z.; Ma, S.; Hu, L.; Hwang, K.; and Liu, Z. 2021. Medical-Level Suicide Risk Analysis: A Novel Standard and Evaluation Model. 8(23): 16825–16834.

Wilson, D.; and Wharton, T. 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.

Xu, Z.; Xu, Y.; Cheung, F.; Cheng, M.; Lung, D.; Law, Y. W.; Chiang, B.; Zhang, Q.; and Yip, P. S. 2021. Detecting Suicide Risk Using Knowledge-Aware Natural Language Processing and Counseling Service Data. 283: 114176.

Zalsman, G.; Levy, Y.; Sommerfeld, E.; Segal, A.; Assa, D.; Ben-Dayan, L.; Valevski, A.; and Mann, J. J. 2021. Suicide-related calls to a national crisis chat hotline service during the COVID-19 pandemic and lockdown. Journal of psychiatric research.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536.