Natural Language Processing (2026), **32**, pp. 357–380
doi:10.1017/nlp.2026.10022

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Combining domain knowledge with large language models for predicting suicide risk in online counselling services

Daniel Izmaylov[1], Avi Segal[1], Kobi Gal[1,2], Meytal Grimland[3], Yossi Levi-Belz[4], Joy Benatov[4] and Yael Levy[5]

[1]Ben-Gurion University of the Negev, Israel, [2]University of Edinburgh, UK, [3]Ruppin Academic Center, Israel, [4]University of Haifa, Israel, and [5]Sahar Online Mental Support
Corresponding author: Kobi Gal; Email: kgal@ed.ac.uk

(Received 11 September 2024; revised 19 January 2026; accepted 16 March 2026; first published online 24 April 2026)

Abstract

Online counselling services have seen increased use in recent years, providing critical emergency mental health support. These interactions are typically long, complex, and varied in the dialogue between help seekers and counsellors. The lack of domain-specific models, especially in low-resource languages, poses a significant challenge for the automatic detection of suicide risk in online chat services for mental health support. To address this challenge, our approach adapts a general-purpose large language model (LLM) to the suicide prediction task that employs a two-stage classification architecture to deal with sparse and imbalanced data. It extends the state of the art by: (1) incorporating psychological theory into model training and (2) capturing key aspects of conversation structure in counselling sessions. We evaluate the performance of the proposed LLM against the state-of-the-art LLMs for suicide detection on thousands of conversations in the Hebrew language from a leading national online counselling service in Israel. Results show that the proposed LLM outperformed existing state-of-the-art approaches in detecting suicide risk, as measured by relevant literature metrics. Moreover, the LLM outperforms other approaches even in the early stages of a conversation, which is crucial for real-time detection in practice. We also discuss the ethical implications of combining LLMs in counselling services. The contributions of this work are (1) extending existing LLM architectures to incorporate domain-specific information; (2) evaluating LLM technologies in the context of socially relevant problems; and (3) introducing novel LLM tools for resource-constrained languages.

Keywords: Evaluation; machine learning; text classification

1. Introduction

Large language models (LLMs) are reshaping natural language processing through their advanced capabilities and enhanced performance. Generic LLMs are trained on large-scale internet data and may perform poorly in tasks requiring domain-specific knowledge, such as text classification or summarisation (Brown *et al.* 2020). As a result, there is increasing research on adapting generic LLMs to domain-specific tasks (Ziegler *et al.* 2019; Hu *et al.* 2023). This paper aims to design and evaluate LLMs for the socially relevant task of suicide prediction. 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 crucial aspect of suicide prevention is the early identification of at-risk individuals (World-Health-Organization, 2021).

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

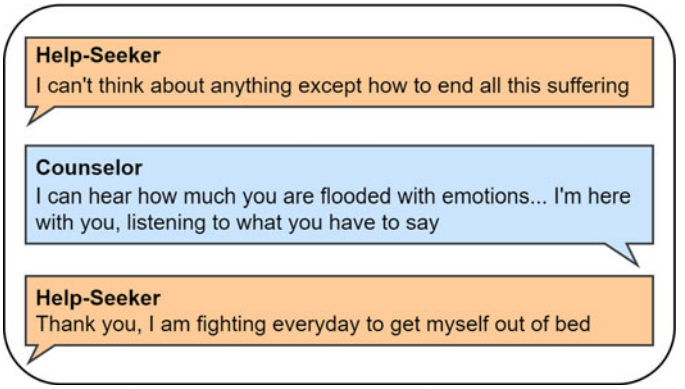


Figure 1. A (fictitious) example of a conversation between a help-seeker and counsellor.

Our approach uses LLMs to detect and assess the severity of suicide risk in anonymous text-based interactions between help-seekers and professional counsellors (see the fictitious example of such an interaction in Figure 1). Our dataset is collected from an online counselling service operating in a low-resource language (Hebrew). Detecting suicide risk in this specific context presents several challenges. These posts differ significantly in structure, length, and context from online conversations between counsellors and help-seekers. Prior research on counselling conversations often overlooks conversation structure or restricts analysis to short exchanges (Cao *et al.* 2019; Bialer *et al.* 2022). Moreover, assessing imminent suicide risk is challenging, even for experienced psychologists, due to its complexity. This complexity is further compounded by data scarcity, as imminent suicide risk is present in less than 2% of conversations in our dataset. Additionally, the NLP resources available for low-resource languages remain significantly limited compared to those for English.

To address this gap, this paper presents SR-BERT, a hierarchical model that includes a base layer for encoding 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 hypothesised that incorporating knowledge from suicide risk theory during the pre-training phase could enhance the detection model's downstream performance. To this end, we develop a new domain knowledge-based pre-training step that embeds a suicide risk factor lexicon (SRF) into SR-BERT. The SRF lexicon was developed by a team of psychologists specialising in suicide risk theory and prevention.

We built on SR-BERT to develop a two-stage model for predicting varying levels of suicide risk severity. This approach not only identifies suicide risk in conversations but also quantifies its severity. Additionally, we applied data augmentation techniques guided by psychological theory to improve model performance.

In empirical studies, we evaluated the capability of different LLMs to predict suicide risk across various dimensions, including global and immediate risk, early detection, and different age groups. We also evaluated the sensitivity of several of these models for several conditions, including role transition and gender. The results generally show that SR-BERT significantly outperforms different classifiers from the literature for suicide risk (SR) detection. Integrating domain-expert information into SR-BERT significantly contributes to its performance.

The contributions of this work are twofold. First in providing a hierarchical model for suicide risk detection in online discussions that combines the structure of the conversation with psychological theory. We show that this approach significantly improves performance in detecting suicide risk from chat conversations compared to the state of the art (Bialer *et al.* (2022)). Second, we expand the set of NLP tools available for resource-constrained languages and are making our code and the SRF lexicon accessible to the broader research community.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

This paper expands on a previous conference version (Izmaylov *et al.* 2023) in two ways. First, it introduces a new model for predicting highly urgent suicide risk, a more challenging problem due to class imbalance and data sparsity. To address this, we propose a two-stage classification model that significantly improves the predictive model's overall performance and enables more precise risk assessment. Second, we provide a sensitivity analysis of our approach, examining its robustness to alterations in conversation structure, role transitions, the emergence of suicide ideation terms, and the help-seeker's declared gender.

2. Related work

Our work relates to past studies in using LLMs to represent domain knowledge and conversation structure in deep language models. 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). Most LLM-based suicide risk studies focus on social media posts (Coppersmith *et al.* 2018; Shing *et al.* 2018; Sawhney *et al.* 2018; Zirikly *et al.* 2019; Tadesse *et al.* 2019; Ophir *et al.* 2020).

Wang *et al.* (2021) combined a generic BERT model with predefined rules for scoring suicide risk in social media. Our work differs by focusing on online counselling conversations, which are longer, structured, and reflect psychological dynamics. We show that modelling these aspects is essential for detecting suicide risk in this context.

Few studies address suicide detection in online counselling and none model the session's conversation structure. The most relevant prior work is by Bialer *et al.* (2022), who combined a BERT-based model (Devlin *et al.* 2018) with a manually curated lexicon of suicide-related terms. This model could only represent a portion of the conversation (512 tokens) and did not account for the counsellor's input. The SR-BERT model extends this approach to reasoning about the entire conversation, making use of the counsellor's input as well as aspects of the conversation structure, such as speaker roles. As we show in the paper, the SR-BERT model significantly outperforms that of Bialer *et al.* (2022) on the same dataset, both for entire conversations and when considering early detection on parts of the conversation.

We mention two SR detection approaches in counselling that did not address early detection. Xu *et al.* (2021) combined a word2vec representation of suicide concepts with a bidirectional LSTM network for SR prediction in Korean online counselling services. 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 require more construction time from human experts. Our approach outperforms a baseline model 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 counselling service; this model did not use a lexicon.

There is strong evidence on the benefits of integrating domain knowledge in language models for downstream tasks (Childs and Washburn 2019; Cao *et al.* 2019; Gaur *et al.* 2019; Lee *et al.* 2020; Colon-Hernandez *et al.* 2021). 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 categories. We extend these approaches by presenting a new method for integrating domain knowledge in the pre-training phase of deep learning models.

In general, NLP models and solutions for low-resource languages are markedly 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. This is the first work to use a hierarchical transformer architecture to model conversation structures in a low-resource language.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

In parallel to suicide risk detection, prior research has examined approaches for identifying positive and supportive language in online interactions. For example, Desmet and Hoste (2013) explored fine-grained emotion detection in suicide notes, including emotions such as hopefulness and thankfulness. Sharma *et al.* (2025) developed an ensemble model to classify hope speech across English and Dravidian languages, while Singh, Sharma, and Singh (2025) introduced EmoGif, a multimodal system to detect emotional support expressed through animated GIFs. These studies demonstrate that combining risk assessment with the detection of positive affect and encouragement can provide a more comprehensive understanding of mental health discourse.

3. The Sahar domain

Sahar (Hebrew acronym for Online Mental Health Support[a]), established in 2000, is Israel's foremost online organisation for emotional support and suicide prevention. It offers anonymous, confidential, and free crisis support through an online chat service available in Hebrew and Arabic. The organisation handles more than 40, 000 chat sessions per year, and this number has significantly increased during the COVID-19 pandemic (Zalsman *et al.* 2021). Due to high demand for the service, counsellors may need to conduct conversations with multiple help seekers in parallel. There is ample research documenting the challenges faced by counsellors in such settings to scale up their support due to increasing traffic (Kitchingman, Caputi, and AlanWoodward 2018; Zabelski *et al.* 2023).

Sahar counsellors are volunteers who receive ongoing guidance and supervision throughout the year from mental health professionals. Counsellors work in shifts, accompanied by trained therapists who monitor conversations and provide professional support as needed. During shifts, counsellors work in a high-stress environment, typically handling multiple chat sessions simultaneously while documenting a summary of each conversation and noting whether the help-seeker shows signs of global suicide risk (GSR). Global suicide risk indicates the presence of suicidal thoughts, without implying immediate or imminent danger. When counsellors determine a help-seeker has an imminent suicide risk (IMSR) – a likelihood of suicidal behaviour in the coming hours or days – they swiftly report the conversation to the authorities. Authorities promptly dispatch a patrol to intervene and prevent harm. Despite their rarity, these cases have a profoundly significant impact. Consequently, an additional tool for early detection and validation of volunteer assessments is necessary to enhance support and intervention.

The Sahar corpus, encompassing a vast collection of over 40, 000 chat sessions spanning a five-year period (2017–2022), serves as a valuable resource for our research. Each conversation within the corpus includes a sequential arrangement of messages exchanged between help-seekers and counsellors, chronologically organised. These conversations serve as the primary data for our analysis and development of the two-stage suicide risk classification model. Approximately 17, 000 of the conversations are labelled by the counsellors.

Table 1 presents general statistics about the dataset. Among the labelled sessions, non-SR conversations constitute the majority, comprising 81.34% of the dataset. These conversations involve individuals seeking help without showing signs of suicide risk. GSR conversations account for 17.67% of the labelled sessions, indicating the presence of general suicidal thoughts. IMSR conversations, constituting only 0.98% of the labelled sessions, represent the most critical cases where immediate intervention is required due to the immediate risk to the individual's life.

GSR conversations are slightly longer and involve more turn exchanges compared to non-SR conversations. This pattern suggests that individuals expressing general suicidal thoughts may require more extensive dialogue to convey their concerns. IMSR conversations, which reflect imminent suicide risk, show an even greater increase in conversation length and turn exchanges, underscoring the urgency and complexity of these cases. Additionally, the mean number of tokens

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

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 1. General statistics for Sahar corpus

Conversation type	Num. of labelled sessions (percentage)	Num. of messages mean (median)	Num. of turn mean (median) exchanges	Num. of tokens mean (median)
Non-SR	14287 (81.34%)	51 (43)	26 (24)	629 (576)
GSR	3104 (17.67%)	57 (49)	30 (28)	676 (616)
IMSR	172 (0.98%)	67 (60)	36 (32)	717 (648)

per conversation rises progressively from non-SR to GSR and IMSR sessions, indicating more detailed and in-depth discussions as suicide risk increases.

To validate the labels, a sample of 500 conversations (200 non SR, 200 GSR, 100 IMSR) was independently labelled by clinical psychologists specialising in suicide. The Krippendorff's α for inter-annotator agreement between the psychologists and the SR label in the conversation is 0.766, which is on 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 this research, a team of psychology experts from a national centre for suicide prevention in Israel has constructed a SRF lexicon in Hebrew that is informed by psychological theory. The SRF lexicon contains terms describing personal and situational factors linked to increased suicidal thoughts, based on validated self-report questionnaires from the psychological and psychiatric literature (Nock *et al.* 2008; Klonsky and May 2015; Turecki and Brent 2016).

Each of the 3,094 sentences in the lexicon was assigned to one of 25 categories. Specifically, terms relating to depression are taken from the Patient Health Questionnaire Depression Module (PHQ-9) (Kroenke, Spitzer, and Williams 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, Steer, and Brown 1996). And terms relating to suicide behaviour 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 burdensome-ness" (translated) included sentences such as "better without me", "I am a burden", and "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.

Similarly, the IMSR lexicon included three layers. The first comprised language representations of key theory-driven factors identified in the suicide literature, specifically the interpersonal theory of suicide (Van Orden *et al.* 2010) and the suicide crisis syndrome (SCS) (Schuck *et al.* 2019). The second layer incorporated validated questionnaires that tap into constructs from key theories (e.g., the Suicide Crisis Inventory for SCS (Galynker *et al.* 2017)). The third layer was developed by reviewing 600 randomly selected chat sessions labelled as suicide risk to identify additional language expressions reflecting these theoretical factors and further enrich the IMSR section of the SRF lexicon.

5. The SR-BERT language model

Our first contribution introduces SR-BERT, a specialised language model designed to identify suicide risk in online conversations. This model predicts suicide risk by combining GSR and IMSR into a single label. SR-BERT extends DialogBERT (Gu, Yoo, and Ha 2021) to better

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

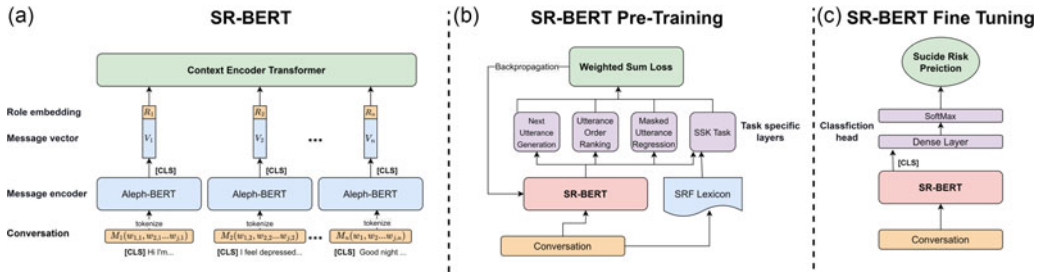


Figure 2. Model architecture. (a) SR-BERT base architecture, encoding conversation and speaker roles. (b) Pre-training procedure on 4 self-supervised tasks including psychological knowledge learning using the SRF lexicon. (c) Fine-tuning procedure learning to predict suicide risk (SR).

capture conversational structure and integrate psychological domain knowledge. The following subsections describe the model architecture and training methodology.

5.1 SR-BERT architecture

The SR-BERT architecture is illustrated in Figure 2(a). It consists of two main components: a transformer encoder for individual messages and a context encoder transformer for modelling the conversation as a whole. The message encoder uses AlephBERT (Seker *et al.* 2022) to convert each message into a vector representation. To capture speaker-specific patterns, this vector is combined with a representation indicating the speaker role (help-seeker or counsellor).

The context encoder transformer operates at the message level rather than the token level. It processes the sequence of encoded message vectors to produce a context-aware representation of the dialogue. This transformer includes 12 attention layers and 12 hidden layers, each with a vector size of 780 to incorporate the speaker role information (compared to 768 in AlephBERT).

This hierarchical design has two benefits. First, it enables modelling turn-taking and speaker dynamics across the conversation. Second, it avoids truncating long dialogues due to the 512-token limit in AlephBERT, since messages are encoded separately.

5.2 Pre-training with self-supervised knowledge

During pre-training (Figure 2(b)), SR-BERT is adapted to counselling data using multiple self-supervised tasks, including a new task based on the SRF lexicon.

For this new task, each conversation is represented as an n-dimensional vector indicating the frequency of sentences belonging to each SRF category. Specifically, the value at index k counts sentences containing at least one term from the kth lexicon category.

We compared a 25-dimensional representation with a reduced 5-dimensional version excluding rare categories (appearing in fewer than 10% of conversations). The five retained categories were "Loss of hope," "Self-injury," "Lack of sense of belonging," "Previous suicide attempt," and "Depression." The reduced representation achieved better performance on validation data using XGBoost.

In this pre-training task, SR-BERT learns to predict the SRF representation from partially masked conversations. Each sentence is randomly masked with an 80% probability. The model then reconstructs the SRF representation from the remaining sentences. Training minimises the mean squared error between the true and predicted vectors. This procedure is repeated with subsets of increasing length to simulate conversations of different durations.

Additionally, we include the three pre-training objectives defined in DialogBERT (Gu, Yoo, and Ha 2021), targeting message semantics, dialogue structure, and sequential order. The final pre-training loss is a weighted sum of the four objectives. Training uses the AdamW optimiser

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

with linear warm-up and an initial learning rate of 5e-5. An adaptive learning-rate scheduler with 0.01 weight decay, 15,000 warm-up steps, and a batch size of 32 is applied. The model is trained for 20 epochs with a fixed random seed. All experiments were conducted on a GeForce RTX 3090 GPU using PyTorch.

5.3 Fine-tuning

In the fine-tuning stage (Figure 2(c)), SR-BERT is adapted for binary suicide risk prediction following the methodology suggested by Sun *et al.* (2020). Specifically, a classification head is added, consisting of a dense layer with output size two and a softmax activation. During fine-tuning, both the context encoder transformer and the classification head are updated to maximise the log-likelihood of the true labels. We use the AdamW optimiser with linear warm-up, an initial learning rate of 2e-5, 0.01 weight decay, and a batch size of 16. The input to the model includes the entire conversation. The model is trained for 10 epochs with a fixed random seed to ensure reproducibility. In the fine-tuning stage (Figure 2(c)), SR-BERT is adapted for binary suicide risk prediction following standard practices reported by Sun *et al.* (2020).

6. Empirical methodology

We randomly split the labelled Sahar dataset to train (70%), validation (15%), and test (15%) sets. The validation set was used to train the model hyperparameters. The test and validation set were excluded from pretraining.

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

- ***SR-BERT w.o. SSK***. This model omits the SSK pre-training task but is otherwise identical to SR-BERT with SSK, including the hierarchical structure and pre-training on the other three tasks.
- **Ensemble SI-BERT** (Bialer *et al.* 2022). Baseline model representing state of the art for suicide detection in online discussions. It was trained on the same Sahar dataset. Ensemble SI-BERT used only the help-seeker text and truncated inputs exceeding this length. We re-implemented the model using the authors' code and parameters and applied it to the dataset used in this study.
- ***SRF-based lexicon + XGBoost***. An XGBoost (Chen and Guestrin 2016) classifier based on the 5-dimension SRF conversation representation over the SRF lexicon. 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 trained on conversation encodings derived from the explicit suicide-related terms proposed by Bialer *et al.* (2022). This list contains 67 terms, including examples such as "commit suicide", "cut wrists", and "wish to die".
- ***Doc2Vec + XGBoost.*** An XGBoost classifier trained on 300-dimensional Doc2Vec representations of each conversation (Le and Mikolov 2014).

7. SR prediction results

We follow prior work in evaluating model performance using ROC-AUC, which is widely employed in suicide detection research (Bernert *et al.* 2020). We also report the F2-score (Sokolova *et al.* 2006) for predicting the positive SR label. This measure assigns greater weight to false negatives, making it appropriate for SR detection, where missing positive cases can have life-threatening consequences.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 2. SR Prediction results of compared models. Bold highlights the 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	69.4	68.3	91.6	69.1	68.8
SR-BERT w. SSK	*78.5*	68.6	*92.1*	*76.3*	*73.2*

7.1 SR detection from complete conversations

Table 2 compares the performance of the SR-BERT model to the baselines in 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. The most notable differences in performance between the SR-BERT w. SSK and the Ensemble SI-BERT model were in the recall metric where SR-BERT w. SSK achieved a 19.1% improvement and 14.0% improvement in the F2 metric.

The SR-BERT w. SSK classifier outperformed the SR-BERT w.o. SSK classifier for all metrics. This supports the hypothesis that incorporating domain knowledge during pre-training can improve SR prediction. We used the McNemar paired test for labelling disagreements (Gillick and Cox 1989) to compare model predictions. Statistical significance with $p < 0.05$ was found for SR-BERT w. SSK *versus* SR-BERT w.o. SSK and for SR-BERT w. SSK *versus* Ensemble SI-BERT.

The SRF lexicon with XGBoost outperformed the explicit lexicon with XGBoost on all metrics except ROC-AUC. The Ensemble SI-BERT achieved the highest precision, slightly outperforming both SR-BERT models. However, it had substantially lower recall, F1, and F2 scores.

We computed precision and recall values for all possible thresholds of the predicted probabilities and plotted these values directly to obtain the precision-recall curves for the SR-BERT w. SSK and SR-BERT w.o. SSK (Figure 3). As seen in the figure, the two models have similar precision at low recall levels. However, for recall above 0.5, the range of primary interest in this study, SR-BERT with SSK shows higher precision.

Finally, to analyse false negatives, we compare the false omission rate across models. This rate is defined as the proportion of false negatives among all instances predicted as negative (see Figure 4). As seen in the figure, the SR-BERT models without SSK reduced the false omission rate compared to Ensemble SI-BERT from 0.083 to 0.071, a 14.4% improvement. Pre-training SR-BERT with the additional SSK task lowered the false omission rate from 0.071 to 0.052, a 26.7% reduction relative to Ensemble SI-BERT.

7.2 Early SR detection

Figure 5 provides important insight into the models' ability to detect suicide risk at different stages of an unfolding conversation. This type of early detection is critical in real-world crisis intervention settings, where prompt identification of risk can inform timely support decisions. By evaluating model performance incrementally across conversation progress, we can assess not only overall accuracy but also how rapidly each model begins to outperform baseline approaches. This contextualises the relative utility of each method when limited information is available, which is often the case in live support environments.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

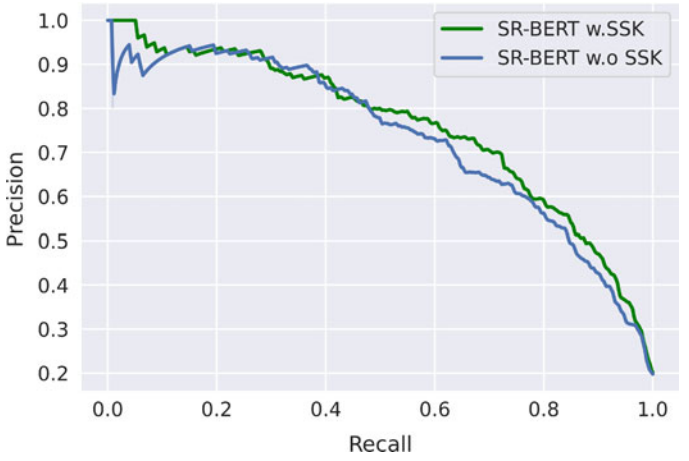


Figure 3. Precision recall curve comparing the results of SR-BERT w. SSK and SR-BERT w.o. SSK.

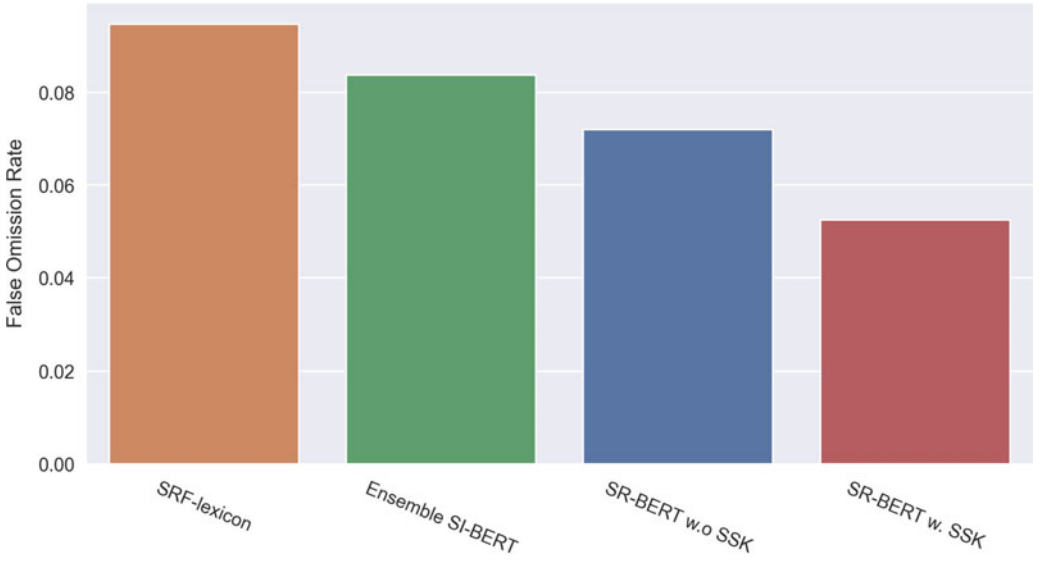


Figure 4. False omission rate of different models.

The length of the conversations varies widely, and longer conversations don't imply harder cases. Therefore, we compare the approaches with respect to their performance after receiving {20, 40, 60, 80, 100} percent of messages in the session, rather than the number of messages. As seen in the figure, all of the models improved in performance as conversations progressed. However, the SR-BERT model with SSK was consistently the top performer, followed by SR-BERT w.o. SSK. The performance gap between the two SR-BERT models grew as the conversation progressed, particularly in the final stages of the conversation (80%–100%). This suggests that the SR-BERT model with SSK better utilises language cues throughout the conversation.

Furthermore, as conversations progress, the performance gap between Ensemble SI-BERT and the SR-BERT models (with and without SSK) widens. One possible explanation is that Ensemble SI-BERT truncates long conversations, potentially omitting crucial information.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

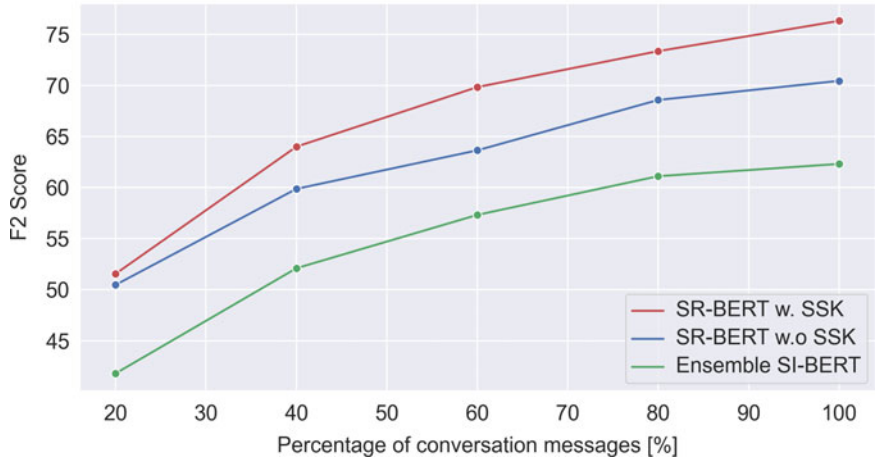


Figure 5. Classification results for early detection of top-performing SR detection approaches.

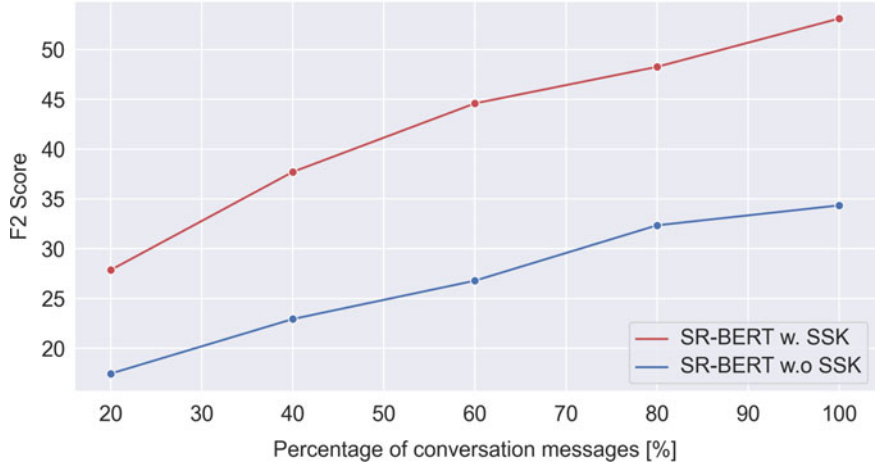


Figure 6. Classification results for early detection on explicit-less-conversations benchmarks.

To further assess the contribution of the SSK pretraining task, we compared the performance of the SR-BERT w. and w.o SSK on a subset of the test set that omits all conversations containing terms from an explicit-ideation-terms lexicon Bialer *et al.* (2022). These conversations do not mention explicit suicide ideation terms and may be more difficult to classify. The performance of SR-BERT with and without SSK pre-training on this dataset is presented in Figure 6. As expected, the overall performance of both models decreased significantly compared to the original test set. However, SR-BERT w. SSK performed better than SR-BERT w.o. SSK and the difference between the two models increased as the sessions progressed. The difference in the performance between the two models was significantly larger than that reported on the original test set (Figure 5). This demonstrates the impact of incorporating domain theory during pre-training to improve SR detection in more challenging cases.

7.3 Sensitivity analysis for the SR model

In this section, we examine the top-performing model, SR-BERT with SSK, and evaluate its sensitivity to different factors: conversation structure, the presence of suicide terms from the lexicon,

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 3. SR-BERT performance for overlap with SRF lexicon (top) and number of tole changes (bottom)

Overlap w. lexicon [%]	*Positive SR [%]*	*ROC-AUC [%]*	*F2*
<1.6	7.77	93.1	58.1
(1.6, 4.5]	12.4	90.2	67.3
(4.2, 8.1]	21.1	89.0	72.7
>8.1	35.4	92.9	85.5
Number of role changes	*Positive SR [%]*	*ROC-AUC [%]*	*F2*
<18.0	13.4	93.7	81.3
(18.0, 24.0]	19.7	91.1	69.5
(24.0, 30.0]	17.7	90.6	72.9
(30.0, 39.0]	19.6	91.5	76.4
>39.0	25.5	93.1	80.3

Table 4. SR-BERT performance for gender subgroups

Gender	*Positive SR [%]*	*F2*	ROC-AUC [%]	*SRF lexicon overlap [%]*
Female (70.5%)	18.5	75.6	92.0	5.9
Male (29.5%)	17.4	76.6	92.2	5.2

and help-seeker gender. To facilitate comparison, we binned conversations into equal-width intervals for each factor (each bin contained the same number of instances).

Table 3 (top) summarises performance relative to SRF lexicon overlap. We observed a moderate Pearson correlation ($r = 0.333$) between overlap and the presence of suicide risk. F2 scores increased with higher overlap, while ROC-AUC scores were slightly higher for conversations with low overlap. This suggests the model relies more on explicit lexicon cues when available.

Table 3 (bottom) shows performance by the number of role changes. We observed a weak Pearson correlation ($r = 0.105$) between more role changes and the presence of suicide risk. This may reflect that counsellors tend to ask more questions when suspecting SR to gather additional information.

Finally, we assessed model performance across help-seeker gender (Table 4). Hebrew's grammatical gender did not lead to systematic differences: F2 and ROC-AUC scores were comparable between male and female help-seekers. The proportion of SRF lexicon overlap was also similar across genders, suggesting no detectable gender bias.

8. Detecting imminent suicide risk

While our previous analyses focused on the binary classification of suicide risk and its early detection, real-world crisis intervention often requires a more fine-grained understanding of risk levels. In particular, distinguishing between general suicidal ideation (GSR) and IMSR is critical for prioritising intervention efforts. Counsellors must act swiftly in IMSR cases, and yet these are rare and difficult to identify.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

8.1 Key challenges

Building on the strength of SR-BERT in SR detection, we now turn to the task of modelling suicide risk severity. By developing a model to assess the severity of suicide risk, we aim to provide counsellors with a more granular understanding of the situations they encounter. This nuanced approach will help them tailor their interventions more effectively.

Moreover, the integration of such a model can serve as a valuable validation instrument for counsellors' evaluations. It can assist in verifying the consistency of severity assessments with the observations made by counsellors, thereby enhancing the overall confidence in the decision-making process.

In our pursuit of accurately predicting IMSR, we are confronted with three primary challenges. Firstly, the task of determining whether a help seeker is at an imminent suicide risk is inherently complex, posing difficulties even for experienced clinical psychologists. Secondly, we face the challenge of limited data availability. Within our dataset, there are only 172 IMSR conversations, which significantly compounds the difficulty of effectively detecting suicidal risk. Lastly, a significant hurdle we must address is the presence of a major class imbalance, as illustrated in Table 1. This class imbalance can adversely affect the performance of the model (Henning *et al.* 2023).

It's worth noting that even within the psychological literature, the accurate identification of high-risk individuals remains a challenging endeavour (Golder *et al.* 2017). It is crucial to underscore that while significant progress has been made in identifying individuals at general suicide risk, differentiating them from those who will actually transition to suicidal behaviour remains an area where substantial work is needed (Singhal, 2020).

To tackle these challenges, we propose a multifaceted approach. Our method for identifying IMSR entails a two-stage model designed to enhance the accuracy of risk assessment. Additionally, we leverage data augmentations by utilising domain knowledge to supplement our dataset. This comprehensive strategy aims to mitigate the inherent complexities posed by the scarcity of data and the intricacies of detecting varying levels of suicidal risk.

8.2 A two-stage model

To address the challenge posed by class imbalance and the inherent complexity of the task, we propose a hierarchy classification method approach. Our strategy involves training a model to focus exclusively on conversations related to SR, distinguishing between GSR and IMSR conversations. We believe that this approach allows the model to better discern the subtle signals that differentiate IMSR from GSR conversations, ultimately leading to improved detection accuracy. This two-stage classification model is illustrated in Figure 7.

The initial phase of our model is dedicated to the binary classification of conversations, specifically to determine whether a conversation is related to suicide (SR) or not (non-suicide-related). SR conversations encompass those labelled as either GSR or IMSR. For the first model, we employ the same model described in the preceding section of this paper.

The second stage of classification focuses on conversations that have been identified as having SR and involves using a more refined model (IMSR-BERT) to differentiate between GSR and IMSR. This model is based on the same pre-trained model as SR-BERT and is fine-tuned specifically on the training examples labelled as GSR or IMSR. The big difference from SR-BERT is that IMSR-BERT does not train on the non-suicide risk instances.

By reducing the overall volume of conversations under consideration and concentrating solely on those labelled as GSR and IMSR, we effectively rebalance the class distribution. This adjustment results in an augmentation of the proportion of IMSR instances within the training dataset, elevating it from a less than 1% representation to a more substantial 5%. This approach enables the model to learn a more nuanced understanding of the different levels of suicide risk present in conversations, ultimately enhancing the performance of our suicide risk detection system.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 5. Prediction results of compared models. Bold highlights the highest value

	Non-SR				GSR				IMSR			
	Recall[%]	Precision [%]	F2	F1	Recall[%]	Precision [%]	F2	F1	Recall[%]	Precision [%]	F2	F1
Flat Model	91.73	93.16	92.01	92.44	68.78	62.67	67.46	65.58	27.45	36.36	19.42	23.53
Two-Stage	84.78	95.73	86.76	89.92	80.98	53.46	73.42	64.4	39.13	42.86	39.82	40.91

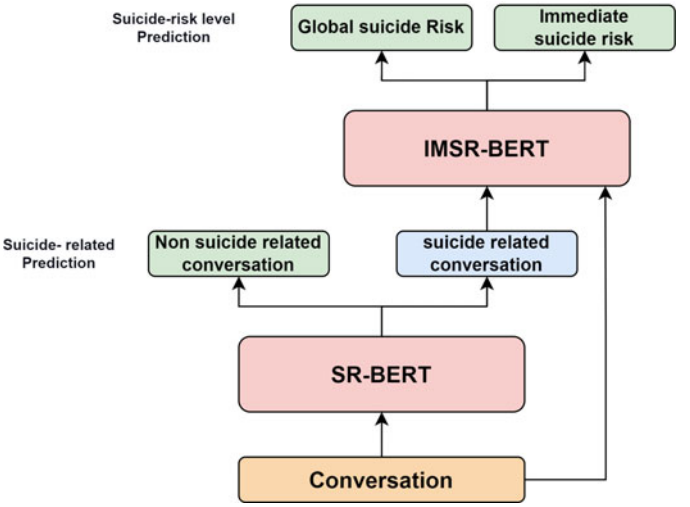


Figure 7. Two stage model for suicide severity classification.

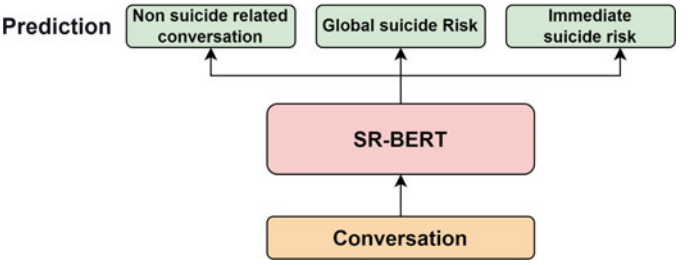


Figure 8. Flat model baseline for suicide severity classification.

9. Two-stage model results

In order to assess the effectiveness of our two-stage model, we conducted a comparative analysis, employing a flat model for reference, as illustrated in Figure 8. The flat model utilised an identical base model as SR-BERT but with a distinct fine-tuning methodology. In contrast to SR-BERT's fine-tuning for binary classification (SR *vs.* non-SR), we fine-tuned the flat model for a three-class prediction task, distinguishing between non-SR, GSR, and IMSR conversations.

The obtained results, presented in Table 5, showcase the performance metrics of both models across various evaluation criteria, using the same metrics as done in Section 7.

From the results, it is evident that the two-stage model outperforms the flat model in most metrics, showcasing its superiority. Specifically, for the IMSR conversations, the two-stage model achieved better results, surpassing the flat model in all evaluated metrics, particularly excelling

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

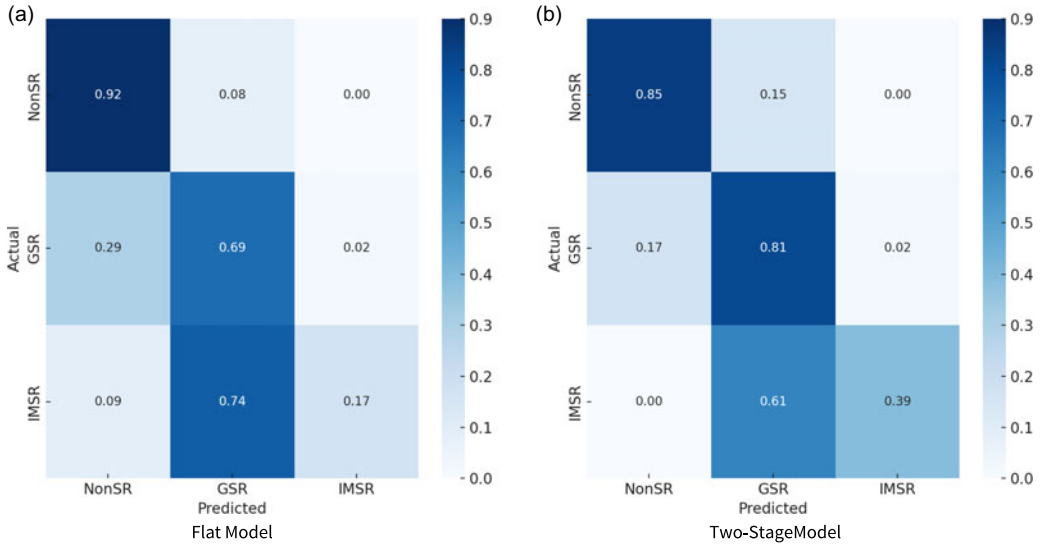


Figure 9. Comparing Models' Confusion Matrices.

in the F2 score. The two-stage model obtained an F2 score of 39.82%, while the flat model's F2 was limited to 19.42%, demonstrating a substantial difference in performance. Additionally, the two-stage model maintained a favourable F1 score of 40.91%, compared to the flat model's 25.53%.

Moving on to the classification of GSR conversations, the two-stage model also displayed notable improvements over the flat model in F2 and recall metrics. It achieved an F2 score of 73.42%, surpassing the flat model's 62.46%. While there was a slight decrease in the F1 score for the two-stage model, reaching 64.4% compared to the flat model's 65.6%, the considerable difference in F2 score highlights the model's improvement in classifying GSR conversations.

Furthermore, we assessed the performance of the two-stage model in recognising non-SR conversations. Here, we observed that the flat model performed slightly better than the two-stage model. To further investigate the implications of this lower performance on non-SR conversations, we conducted an analysis of the confusion matrices for both models (Figure 9).

Upon analysing the confusion matrix of the flat model, as depicted in Figure 9a, we observed that the model demonstrated accurate classification for non-SR conversations, with relatively few mistakes. However, the model struggled to correctly classify IMSR conversations, mislabelling 9% of the conversations as non-SR. Additionally, around 30% of the GSR conversations were erroneously labelled as non-SR, indicating the need to minimise such misclassifications for both use cases.

In comparison, the confusion matrix of the two-stage model, shown in Figure 9b, revealed its improved ability to classify GSR conversations accurately, without mistakenly labelling any IMSR conversations as non-SR. Despite this advancement, there still exists room for improvement, as several IMSR conversations were not correctly classified.

Overall, the results underscore the strengths of the two-stage model in effectively distinguishing between different conversation classes, particularly exhibiting significant improvements in IMSR and GSR classification. While the flat model exhibited a slight advantage in non-SR conversation identification, the two-stage model's superior performance across other crucial classes renders it a promising choice for our specific task. To further enhance the model's performance, we explore data augmentation techniques.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

10. Data augmentation

To improve the performance of our second-stage model, we hypothesise that the relatively low performance in IMSR prediction may stem from the limited sample size associated with this label and the overall class imbalance in the dataset. This hypothesis is consistent with prior research findings (Feng *et al.* 2021; Henning *et al.* 2023) that indicate how limited sample sizes can have a detrimental impact on model performance.

10.1 Proposed methods

To address this challenge, we have undertaken a comprehensive exploration of various sampling strategies designed to increase the representation of IMSR conversations within our training dataset.

Under sampling We randomly sampled and removed a certain percentage (x%) of the GSR conversations to reduce their dominance in the training data.

Over sampling We randomly resampled the IMSR conversations to amplify their presence in the training dataset. Doing so ensures that the model receives more exposure to IMSR conversations, which are relatively rarer in the original dataset.

Lexicon-Based Augmentation To enrich the existing conversations, we utilised the IMSR lexicon. Our data augmentation approach involved substituting terms within the conversations using the lexicon. Specifically, we replaced all the terms that are present in the IMSR-SRF-factors lexicons with other terms that belong to the same SRF factors. By introducing these targeted substitutions, we introduced small variations to the text, thereby enabling the model to learn from a broader range of linguistic patterns and contexts related to suicide risk. For example, in this message, "The rope is waiting for me", the category is "Active Suicidal Ideation with Specific Plan and Intent" and the message can be changed to "The gun is ready for me".

In order to evaluate the model's performance, we conducted a comparative analysis of the two-stage model trained on the same dataset with varying rates of oversampling and undersampling, ranging from 1 to 5 and 10. Oversampling was performed for the minority class (IMSR), and undersampling was performed for the majority class (non-IMSR). Additionally, we compared the performance of normal oversampling to that of using "Lexicon Based Augmentation" for each oversampling rate. Through these experiments with different sampling techniques, our goal was to understand how these adjustments impact the model's effectiveness in classifying conversation types.

10.2 Data augmentation results

Table 6 summarises the data augmentation trails. As seen by the table, among the different approaches, the "undersample_20%" model, which employs undersampling with a ratio of 5, demonstrated competitive results. It achieved a recall of 79.51% and an F2 score of 72.35% for GSR conversations, slightly under-performing compared to the model trained on the untreated dataset. However, it significantly improved its performance in classifying IMSR conversations, achieving a recall of 52.17% and an F2 score of 49.59%.

On the other hand, the "oversample_500%" model, which utilises oversampling with a ratio of 5, exhibited notable improvements in GSR classification, attaining a recall of 82.68% and a precision of 53.72%. However, its performance in IMSR classification was comparatively lower than the untreated model in terms of recall and F2. We investigated the impact of "Lexicon Based Augmentation" in conjunction with oversampling, as demonstrated by the "oversample_500%_aug" model. Although this approach improved the precision for IMSR conversations to 70.01%, the recall remained relatively low at 30.43%. Finally,

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 6. Prediction results of compared models. Bold highlights the highest value

	GSR				IMSR			
Model	Recall[%]	Precision [%]	F2	F1	Recall[%]	Precision [%]	F2	F1
Two- Stage	80.98	53.46	73.42	64.4	39.13	42.86	39.82	40.91
undersample_20%	79.51	53.18	72.35	63.73	52.17	41.38	49.59	46.15
oversample_500%	82.68	53.72	74.64	65.13	30.4	63.64	33.98	41.18
oversample_500%_aug	82.44	53.48	74.38	64.88	30.43	70.01	34.31	42.42
oversample_500%_undersample_20%	84.78	53.71	72.54	64.11	52.17	34.29	47.24	41.38
oversample_500%_undersample_20%_aug	79.02	53.38	72.1	63.72	60.87	40.04	55.12	48.28

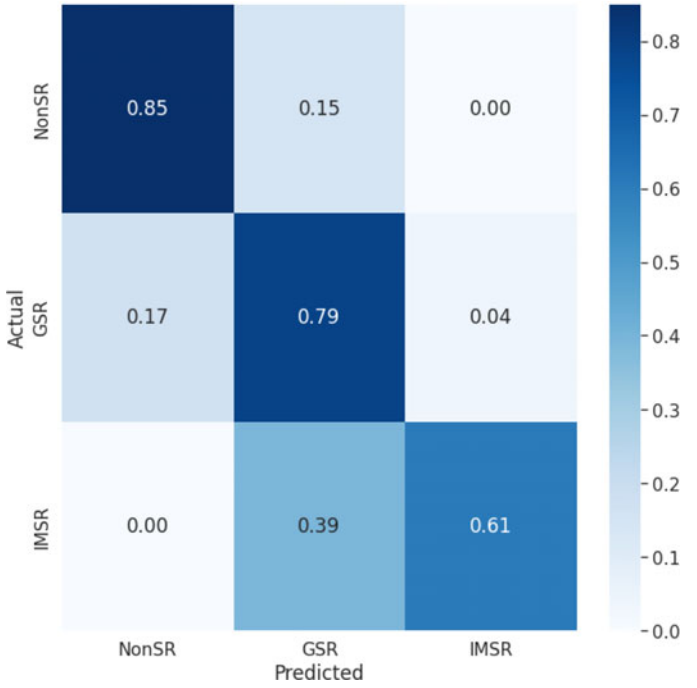


Figure 10. Confusion matrix for the data-augmented model.

the "oversample_500%_undersample_20%_aug" method, which combined oversamping and undersanpling, performed the best regarding the recall and F2 score for IMSR, achieving 60.87% and 55.12%, respectively. This indicates the effectiveness of combining oversampling, undersampling, and data augmentation in improving IMSR classification while performing only slightly worse than the model trained on the untreated training set.

Another noteworthy aspect to consider is that due to the architecture of the two-stage model, a higher error rate in the classification of GSR conversations may result in their erroneous labelling as IMSR conversations. While this does introduce a relatively low-risk misclassification scenario, we conducted an analysis employing confusion matrix evaluation to delve deeper into this phenomenon.

Figure 10 presents the confusion matrix for the data-augmented model. As anticipated, the performance of the non-SR class remained consistent compared to the confusion matrix shown

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

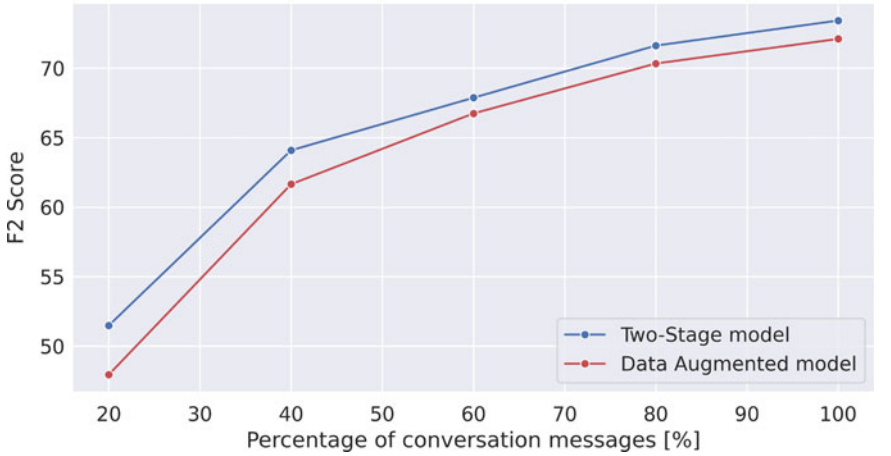


Figure 11. Classification results on GSR class for early detection.

in ?? (non-augmented model). In contrast, there is a notable improvement in the recall of IMSR conversations. However, it is worth noting that approximately 2% more GSR conversations were incorrectly labelled as IMSR when compared to the non-augmented two-stage model. This particular misclassification is relatively minor in our context because we can tolerate predictions that lean towards a more severe risk prediction. Given our model's architecture, conversations initially labelled as SR can only be classified as either GSR or IMSR in subsequent steps.

Overall, our experiments with different sampling techniques shed light on their impact on the model's performance in classifying conversation types. The findings suggest that a careful combination of sampling methods with data augmentation can lead to significant improvements in IMSR classification without compromising the performance of GSR conversations. These insights provide valuable guidance for optimising models' performance in real-world NLP applications.

11. Early detection of two stage models

We employed the same evaluation methodology as conducted in Subsection 7.2 to assess the performance differences in early detection. From Figure 11, we can observe the performance of the model's for GSR conversations, where both models demonstrated similar behaviour. Notably, we observed only marginal performance reductions in the data-augmented model compared to the two-stage model. As anticipated, the performance of both models improved as the conversations progressed. This consistent pattern persisted across various percentages of the conversation duration.

Conversely, from Figure 12 we can notice a substantial performance enhancement emerged for the IMSR's F2 metric. The augmented model achieved superior results compared to the two-stage model, particularly exhibiting noteworthy performance (F2 score of 39.06) even in the early stages of conversations. This starkly contrasts with the two-stage model's performance (F2 score of 5.15) at the same early juncture. Moreover, the performance improvement in the data-augmented model exhibited only minor fluctuations as the conversations unfolded, unlike the pronounced performance boost observed in the two-stage model. This intriguingly suggests that the data-augmented model can effectively identify IMSR conversations at an early stage, potentially indicating the presence of IMSR even in the initial phases of conversations. This finding implies that certain cues or indications of IMSR might be detectable early in a conversation, contributing to the data-augmented model's enhanced early detection performance.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 7. Two-Stage model performance for gender subgroups

Gender	GSR [%]	F2 - GSR	IMSR [%]	F2-IMSR	IMSR-Lex Overlap [%]
Female (70.5%)	18.5	71.02	0.94	47.61	1.50
Male (29.5%)	17.4	74.66	1.24	69.76	1.52

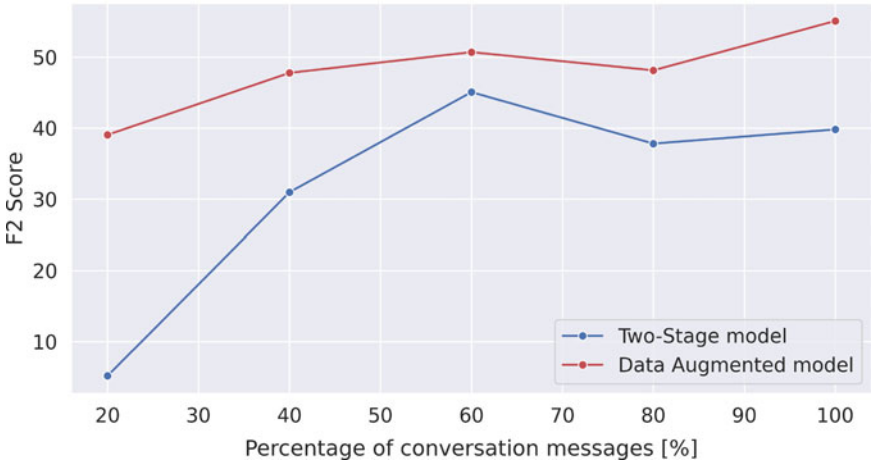


Figure 12. Classification results on IMSR class for early detection.

11.1 Sensitivity analysis – two-stage model

From the results depicted in Table 7, it is apparent that, despite an almost twofold representation of female help-seekers, the numbers of GSR conversations from both genders are roughly equal. Additionally, a noteworthy discovery is the higher frequency of IMSR conversations initiated by male help-seekers. This observation raises the possibility that when a male help-seeker engages in a conversation, it may indicate a more severe mental state. Furthermore, it is evident that the percentage of IMSR-lexicon overlap is smaller compared to that of GSR; however, the degree of overlap between female and male help-seekers is quite similar.

Comparing these findings with the outcomes in Table 4, it becomes evident that the two-stage model exhibits a susceptibility to bias, manifested in its F2 scores for both GSR and IMSR conversations. This suggests that gender-related disparities are affecting the model's performance, potentially leading to biased predictions in conversations initiated by different genders. Such biases can significantly impact the model's practical applicability, underscoring the importance of addressing these issues in order to enhance its effectiveness across diverse user demographics.

12. Discussion

Our results demonstrate the importance of combining domain knowledge and conversation structure with pre-trained language models for the purpose of GSR and IMSR detection. In particular, embedding domain theory into the SR-BERT pre-training step captures nuances of the conversations that go beyond explicit mentions of suicide-related terms. Furthermore, the hierarchical SR-BERT approach was able to overcome the problem of limited input size that hinders existing approaches to language modelling. It was able to identify suicide risk variables early

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Table 8. Illustrative excerpts from conversations where SR-BERT and Ensemble SI-BERT disagree

Help Seeker:	"I generally feel it's over for me. "
Counsellor:	"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. "
Counsellor:	"If I understood correctly, you slapped your sister and locked yourself in the toilet. "
Help Seeker:	"Correct. "
Counsellor:	"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. "

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, in actual support sessions.

We demonstrate the SR model's ability to capture nuanced textual exchanges with the two illustrative examples presented in Table 8. 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 counsellor 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 hypothesise that the flat ensemble model was misled by the "gun", "police", and "trauma" utterances while SR-BERT was able to weigh these terms in context.

We also looked at the samples that the models failed to predict correctly to better define future needed improvements. For this, we have enlisted a domain expert who analysed the errors performed by the model. Specifically, this expert analysed 50 GSR conversations and 20 IMSR conversations that were misclassified. The expert's analysis revealed that misclassifications may occur in GSR calls when many topics are raised during the conversation or when the help-seeker is "beating around the bush" for a long time with only very gently referencing their mental states. In IMSR calls the expert pointed mainly to misclassifications due to short calls that were abruptly disconnected by the help-seeker. We hypothesise that improving the taxonomy used by the model, incorporating explicit time-based cues in the input data (such as utterance duration) and moving to bigger language models which capture larger contexts may help in addressing some of these failings.

13. Conclusion and future work

In conclusion, this study has introduced a novel automated approach for evaluating the severity of suicide risk within online interactions involving individuals seeking help and professional counsellors. Early identification of individuals at risk is a fundamental objective in suicide prevention efforts.

Our approach extends the state-of-the-art in deep language modelling 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 counsellors, and (3) adapting to a low-resource language (Hebrew).

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Furthermore, we have shown that the model can not only predict the presence of suicide risk but also assess its severity using a two-stage architecture enriched by domain-informed data augmentation techniques. This approach improves the ability to identify and address suicide risk in online conversations.

The presented approach was able to significantly outperform the state-of-the-art approaches when detecting multiple levels of suicide risk from complete conversations, as well as early detection when only part of the conversation is available.

For future work, we intend to improve our approach by capturing more aspects of conversations, such as prosody (Wilson and Wharton 2006; Kliper, Yonatan, and Daphna 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.

Additionally, we aim to explore how our model enhances volunteers' judgement in suicide risk assessment. This involves investigating its impact on decision-making processes within real-world counselling settings. By conducting empirical studies, we intend to evaluate the model's practical utility and effectiveness. These insights will be critical for understanding its influence on volunteer assessments and identifying potential areas for improvement, ultimately contributing to better outcomes in suicide prevention efforts.

Finally, an important avenue for future research is to complement risk detection models with systems that can also recognise and promote supportive or hopeful content. Detecting encouragement, positive affirmations, and emotional support could help counterbalance distress signals and contribute to more holistic interventions. Prior work on hope speech detection (Sharma *et al.* 2025) and multimodal analysis of supportive messages (Singh, Sharma, and Singh, 2025) highlights the feasibility and potential impact of integrating positivity-focused approaches alongside risk assessment tools.

14. Limitations

We note several limitations of this study.

First, our model was only evaluated in the Hebrew language. We have not directly compared our models 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. Nonetheless, we note that the methodology presented in this manuscript is generalisable to other languages and cultural contexts. Adapting the approach to a new language involves two main components: (1) identifying or training a suitable pre-trained language model in the target language (e.g., AraBERT for Arabic); and (2) constructing or adapting a domain-specific lexicon reflecting suicide risk factors in that language and culture. To ensure the theoretical grounding and cultural relevance of such lexicons, we recommend involving local mental health professionals to translate, validate, or create terms based on the same underlying psychological theories (e.g., PHQ-9, INQ, and SCS).

In fact, as part of our ongoing work, we are developing an Arabic version of SR-BERT using a translated version of the SRF lexicon curated by clinical psychologists fluent in Arabic and familiar with the relevant cultural idioms of distress. We also leverage AraBERT (Antoun, Baly, and Hajj 2020) as the base transformer. This demonstrates the practical feasibility of adapting our architecture across languages. Future work could explore cross-lingual pretraining or multilingual foundation models to further simplify this transfer.

Second, the annotation of the help seekers' mental state was performed by the counsellors, rather than the help seekers themselves. Although the counsellors completed several months of training and were supervised by certified clinical psychologists, it is still possible that they misclassified the help-seekers' mental states. This issue is prevalent in many studies that rely on observer-reported data. Specifically, while counsellor annotations provide valuable real-world

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

labels, they may introduce systematic biases. For example, counsellors' assessments can be influenced by personal clinical experience, training differences, or cognitive heuristics (e.g., confirmation bias). This could affect the consistency and objectivity of suicide risk labels, especially in nuanced or ambiguous cases. Additionally, since counsellors multitask and work under emotional stress, subtle signs may be overlooked or overemphasised.

As an alternative, annotation by panels of trained clinical psychologists working independently or via consensus can improve label consistency. Another approach would be to incorporate help-seeker self-assessment surveys post-conversation (though impractical in many real-world settings). In our case, we partially mitigated this concern by validating a subset of the dataset (500 conversations) with independent expert psychologists, achieving a Krippendorff's α of 0.766, which is comparable to similar work in the literature. Future work could explore these ensemble annotation approaches to better help-seekers' mental states.

Third, our experiments used a random-based train-test split (as opposed to time-based splits) and did not use cross validation. We chose random splits to ensure broad coverage of conversation variability and avoid confounds from external events. Time-based splits in our dataset would have introduced biases due to disruptions like COVID-19 and other crises, which caused abrupt shifts in conversation content and volume. Moreover, our training pipeline involves costly pretraining and augmentation steps, making repeated time-based training or cross-validation infeasible. Nonetheless, temporal robustness is important, and this is a direction for future work.

Fourth, we did not use generative models (such as ChatGPT) in this research due to privacy considerations and since such models were not yet available for the Hebrew language in an open-source configuration (which does not require data sharing).

Finally, while our current model emphasises predictive performance and the integration of domain-specific psychological knowledge, we recognise the importance of model interpretability in sensitive applications such as suicide risk detection. As a step toward addressing this critical concern, we have conducted a complementary study exploring the "black box" of suicide risk prediction models. In this work (Grimland *et al.* 2024), we analysed the contribution of specific psychological risk factors (e.g., hopelessness, thwarted belongingness) to model decisions by leveraging a theory-driven lexicon and logistic regression analysis. This enabled a more transparent understanding of which features and linguistic cues are associated with increased suicide risk in real-time conversations. We view these efforts as part of a broader agenda toward explainable AI in crisis intervention, and our future work will integrate such insights directly into real-time, interpretable models.

Ethics statement

This study was conducted in accordance with ethical standards and approved by the institutional review boards of the participating institutions. All data used in this study, including the Sahar conversation corpus and the SRF psychological lexicon, were obtained in compliance with IRB requirements. Specifically, the Sahar dataset was anonymised and encrypted to protect participant privacy. All help-seekers provided informed consent for the anonymous use of their sessions in research. The counsellors signed consent forms allowing their text data to be used for the study.

Although the model can successfully predict GSR and IMSR during conversations, it is not intended to replace human volunteer counsellors. We believe human involvement is essential in supporting help-seekers. The automated model is intended to assist counsellors by enhancing their ability to assess conversation severity, not to replace them. We believe that in the future, once all necessary approvals and adaptations are in place, such models could be deployed as supportive tools. Rather than replacing counsellors, they would serve as an additional source of insight, like a 'friendly parrot', to assist decision-making in the high-pressure situations counsellors face on a daily basis.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Acknowledgements. This research was funded in part by Israeli Science Foundation grant no. 1302/21.

References

Antoun W., **Baly F.** and **Hajj H.** (2020). Arabert: transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, With a Shared Task on Offensive Language Detection*, pp. 9–15.

Bantilan N., **Malgaroli M.**, **Ray B.** and **Hull T.D.** (2021). Just in time crisis response: Suicide alert system for telemedicine psychotherapy settings. *Psychotherapy Research: Journal of the Society for Psychotherapy Research* **31**(3), 302–312. https://doi.org/10.1080/10503307.2020.1781952.

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 & 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. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4241–4250.

Brown T., **Mann B.**, **Ryder N.**, **Subbiah M.**, **Kaplan J.D.**, **Dhariwal P.**, **Neelakantan A.**, **Shyam P.**, **Sastry G.**, **Askell**, **A.**, **Agarwal**, **S.**, **Herbert-Voss**, **A.**, **Krueger**, **G.**, **Henighan**, **T.**, **Child**, **R.**, **Ramesh**, **A.**, **Ziegler**, **D. M.**, **Wu**, **J.**, **Winter**, **C.**, **Hesse**, **C.**, **Chen**, **M.**, **Sigler**, **E.**, **Litwin**, **M.**, **Gray**, **S.**, **Chess**, **B.**, **Clark**, **J.**, **Berner**, **C.**, **McCandlish**, **S.**, **Radford**, **A.**, **Sutskever**, **I.** and **Amodei**, **D.** (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* **33**, 1877–1901.

Cao L., **Zhang H.**, **Feng L.**, **Wei Z.**, **Wang X.**, **Li N.** and **He X.** (2019). Latent Suicide Risk Detection on Microblog via Suicide-OrientedWord Embeddings and Layered Attention. October 26, 2019. arXiv: 1910.12038 [cs]. http://arxiv.org/abs/1910.12038 (Accessed 15 November 2021).

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, pp. 785–794.

Childs C.M. and **Washburn N.R.** (2019). Embedding domain knowledge for machine learning of complex material systems. *MRS Communications* **9**(3), 806–820. https://doi.org/10.1557/mrc.2019.90. issn: 2159-6859, 2159-6867, (accessed 14 December 2021). http://link.springer.com/10.1557/mrc.2019.90.

Chriqui A. and **Yahav I.** (2021). HeBERT & HebEMO: A Hebrew BERT Model and a Tool for Polarity Analysis and Emotion Recognition. February 25, 2021. arXiv: 2102.01909 [cs]. http://arxiv.org/abs/2102.01909 (Accessed 27 November 2021).

Colon-Hernandez P., **Havasi C.**, **Alonso J.**, **Huggins M.** and **Breazeal C.** (2021). Combining pre-trained language models and structured knowledge. February 5, 2021. arXiv: 2101.12294 [cs]. 2021. http://arxiv.org/abs/2101.12294 (Accessed 5 December 2021).

Coppersmith G., **Leary R.**, **Crutchley P.** and **Fine A.** (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights* **10**, 1178222618792860.

Desmet B. and **Hoste V.** (2013). Emotion detection in suicide notes. *Expert Systems with Applications* **40**(16), 6351–6358.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pp. 4171–4186.

Feng S.Y., **Gangal V.**, **JasonWei S.C.**, **Vosoughi S.**, **Mitamura T.** and **Hovy E.** (2021). A survey of data augmentation approaches for NLP. no. arXiv:2105.03075 (December 1, 2021). arXiv: 2105.03075 [cs]. 2021-12-01. http://arxiv.org/abs/2105.03075 (Accessed 5 June 2023).

Galynker I., **Yaseen Z.S.**, **Cohen A.**, **Benhamou O.**, **Hawes M.** and **Briggs J.** (2017). Prediction of suicidal behavior in high risk psychiatric patients using an assessment of acute suicidal state: The suicide crisis inventory. *Depression & Anxiety* **34**(2), 147–158.

Gaur M., **Amanuel A.**, **Joy P.S.**, **Ugur K.**, **Krishnaprasad T.**, **Ramakanth K.**, **Amit S.**, **Randy W.** and **Jyotishman P.** (2019). Knowledge-aware assessment of severity of suicide risk for early intervention. In The WorldWideWeb Conference on -WWW'19. San Francisco, CA, USA: ACM Press, pp. 514–525. https://doi.org/10.1145/3308558.3313698 (accessed 5 December 2021).

Gillick L. and **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.

Golder V., **Kandane-Rathnayake R.**, **Hoi A.Y.-B.**, **Huq M.**, **Louthrenoo W.**, **An Y.**, **Li Z.G.**, **Luo S.F.**, **Sockalingam S.** and **Lau C.S.** (2017). Association of the lupus low disease activity state (LLDAS) with health-related quality of life in a multinational prospective study. *Arthritis Research & Therapy* **19**(1), 1–11.

Grimland M., **Benatov J.**, **Yeshayahu H.**, **Izmaylov D.**, **Segal A.**, **Gal K.** and **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**(3), 416–424.

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

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, vol. **35**, pp. 12911–12919.

Henning S., **Beluch W.**, **Fraser A.** and **Friedrich A.** (2023). A survey of methods for addressing class imbalance in deep-learning based natural language processing. arXiv:2210.04675, February 22, 2023. arXiv: http://arxiv.org/abs/2210.04675 [cs]. http://arxiv.org/abs/2210.04675 (Accessed 5 June 2023).

Hu L., **Liu Z.**, **Zhao Z.**, **Hou L.**, **Nie L.** and **Li J.** (2023). A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering* **36**(4), 1413–1430.

Izmaylov D., **Segal A.**, **Gal K.**, **Grimland M.** and **Levi-Belz Y.** (2023). Combining psychological theory with language models for suicide risk detection. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2385–2393.

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. *IEEE Transactions on Computational Social Systems* **8**(1), 214–226. https://doi.org/10.1109/TCSS.2020.3021467.

Kitchingman T.A., **Caputi P.**, **Woodward A.**, **Wilson C.J.**, **Wilson I.** and **Soundy A.** (2018). The impact of their role on telephone crisis support workers' psychological wellbeing and functioning: Quantitative findings from a mixed methods investigation. *PLoS one* **13**(12), e0207645.

Kliper R., **Yonatan V.**, **Daphna W.** and **Shirley P.** (2010). Evidence for depression and schizophrenia in speech prosody. In *Proceedings of the 3rd ICSA Tutorial and Research Workshop on Experimental Linguistics 2010*, pp. 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.W.** (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. In Eric P.X. and Tony J. (eds), *Proceedings of Machine Learning Research*, vol. **32**, no. 2 (22-24 Jun), pp. 1188–1196. https://proceedings.mlr.press/v32/le14.html.

Lee D., **Park S.**, **Kang J.**, **Choi D.** and **Han J.** (2020). Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 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. *Scientific Reports* **10**(1), 16685. https://doi.org/10.1038/s41598-020-73917-0 (accessed 15 November 2021).

Posner K., **Brent D.A.**, **Lucas C.**, **Gould M.**, **Stanley B.**, **Brown G.K.**, **Fisher P.**, **Zelazny J.**, **Burke A.**, **Oquendo M.** and **Mann J.J.** (2008). *Columbia-Suicide Severity Rating Scale (C-SSRS)*. New York, NY: Columbia University Medical Center.

Sawhney R., **Manchanda P.**, **Mathur P.**, **Shah R.** and **Singh R.** (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*, pp. 167–175.

Schuck A., **Calati R.**, **Barzilay S.**, **Bloch-Elkouby S.** and **Galynker I.** (2019). Suicide crisis syndrome: A review of supporting evidence for a new suicide-specific diagnosis. *Behavioral Sciences & the Law* **37**(3), 223–239.

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)*, pp. 46–56.

Sharma D., **Gupta V.**, **Singh V.K.** and **Chakravarthi B.R.** (2025). Stop the hate, spread the hope: An ensemble model for hope speech detection in English and Dravidian languages. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Shing H.-C., **Nair S.**, **Zirikly A.**, **Friedenberg M.**, **Daumé H.**III and **Resnik P.** (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*, pp. 25–36.

Singh A., **Sharma D.** and **Singh V.K.** (2025). EmoGif: A multimodal approach to detect emotional support in animated GIFs. *IEEE Transactions on Computational Social Systems* **12**(5), 3791–3803.

Singhal T. (2020). A review of coronavirus disease-2019 (COVID-19). *The Indian Journal of Pediatrics* **87**(4), 281–286.

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*. Springer, pp. 1015–1021.

Sun C., **Qiu X.**, **Xu Y.** and **Huang X.** (2020). How to fine-tune BERT for text classification? (February 5, 2020). arXiv: 1905.05583 [cs]. http://arxiv.org/abs/1905.05583 (Accessed 6 April 2022).

Tadesse M.M., **Lin H.**, **Xu B.** and **Yang L.** (2019). Detection of suicide ideation in social media forums using deep learning. *Algorithms* **13**(1), 7.

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

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press

Van Orden K.A., **Cukrowicz K.C.**, **Witte T.K.** and **Joiner T.E.** (2012). Thwarted belongingness and perceived burdensomeness: Construct validity and psychometric properties of the Interpersonal Needs Questionnaire. *Psychological Assessment* **24**(1), 197–215.

Van Orden K.A., **Witte T.K.**, **Cukrowicz K.C.**, **Braithwaite S.R.**, **Selby E.A.** and **Joiner T.E.** (2010). The interpersonal theory of suicide. *Psychological Review* **117**(2), 575–600.

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. *IEEE Internet of Things Journal* **8**(23), 16825–16834. https://doi.org/10.1109/JIOT.2021.3052363.

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*. World Health Organization.

Xu Z., **Xu Y.**, **Cheung F.**, **Cheng M.**, **Lung D.**, **Law Y.W.**, **Chiang B.**, **Zhang Q.** and **Yip P.S.F.** (2021). Detecting suicide risk using knowledge-aware natural language processing and counseling service data. *Social Science & Medicine* **283**, 114176. https://doi.org/10.1016/j.socscimed.2021.114176 (accessed 19 November 2021).

Zabelski S., **Kaniuka A.R.**, **A. Robertson R.** and **Cramer R.J.** (2023). Crisis lines: Current status and recommendations for research and policy. *Psychiatric Services* **74**(5), 505–512.

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* **139**, 193–196.

Ziegler D.M., **Stiennon N.**, **Wu J.**, **Brown T.B.**, **Radford A.**, **Amodei D.**, **Christiano P.** and **Irving G.** (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv: 1909.08593.

Zirikly A., **Resnik P.**, **Uzuner O.** and **Hollingshead K.** (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*, pp. 24–33.

Cite this article: Izmaylov D, Segal A, Gal K, Grimland M, Levi-Belz Y, Benatov J and Levy Y (2026). Combining domain knowledge with large language models for predicting suicide risk in online counselling services. *Natural Language Processing* **32**, 357–380. https://doi.org/10.1017/nlp.2026.10022

https://doi.org/10.1017/nlp.2026.10022 Published online by Cambridge University Press