Detecting Suicide Risk in Online Counseling Services: A Study in a Low-Resource Language

Amir Bialer[†] **Daniel Izmaylov**[†] **Avi Segal**[†] **Oren Tsur**[†]

{amirbial,zmaylov}@post.bgu.ac.il avisegal@gmail.com orentsur@bgu.ac.il

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

Kobi Gal[†§]
kobig@bgu.ac.il

[†]Ben-Gurion University of the Negev [‡]Ruppin Academic Center [§]University of Edinburgh

Abstract

With the increased awareness of situations of mental crisis and their societal impact, online services providing emergency support are becoming commonplace in many countries. Computational models, trained on discussions between help-seekers and providers, can support suicide prevention by identifying at-risk individuals. However, the lack of domain-specific models, especially in low-resource languages, poses a significant challenge for the automatic detection of suicide risk. We propose a model that combines pre-trained language models (PLM) with a fixed set of manually crafted (and clinically approved) set of suicidal cues, followed by a two-stage fine-tuning process. Our model achieves 0.91 ROC-AUC and an F2-score of 0.55, significantly outperforming an array of strong baselines even early on in the conversation, which is critical for real-time detection in the field. Moreover, the model performs well across genders and age groups.

1 Introduction

The World Health Organization (WHO) lists suicide as one of the most salient causes of death world-wide, causing more deaths than breast cancer or war (WHO, 2019). With close to 1 million lives taken directly by suicide every year, and over 25 million suicide attempts, suicide incurs a lasting impact on families and communities. Identifying individuals at risk ahead of time and providing them with psychological and medical support is a key step in suicide prevention (Joiner et al., 2018).

In this paper, we tackle the task of detecting Suicide Ideation (SI). Specifically, we aim at the detection of SI of individuals contacting online counseling hot-lines in low-resource and morphologically-rich languages. We focus on anonymous data from online, text-based (chat), support services. Such services are available in many countries, allowing for confidential and immediate help to those in distress, and play a critical role in suicide prevention (Bantilan et al., 2021; Jashinsky et al., 2014; Joiner et al., 2007). Empirical evidence suggests that at-risk individuals seek help in close proximity to actual suicide attempts (Zalsman et al., 2021), thus it is critical to identify suicide risks as early as possible during the session.

There is a growing body of work on assessing suicide risk from English texts (whether in social media posts or from counseling sessions), but there is an acute lack of NLP resources that could be used for detection of suicide risk in other languages (Lee et al., 2020). We directly address this gap by focusing on suicide risk detection from online counseling services in Hebrew, which is a low-resource and morphologically-rich language that challenges traditional NLP tasks (Seker et al., 2021).

Our proposed approach for the SI detection task, called SI-BERT, extends a generic pretrained model with a small set of Out of Vocabulary (OOV) tokens, pretraining the language model on a masked LM task and fine tuning for the SI classification task. We further train a logistic regression model, based on a manually crafted lexicon of suicide ideation terms (vetted by domain experts). We then create an Ensemble model by training SI-BERT together with the lexicon. Both SI-BERT and the Ensemble model outperform alternative approaches ranging from W2V (Word2Vec) embeddings and feed-forward networks to Hebrew psychological lexicon. Additionally, the Ensemble model achieves 82% ROC-AUC even when processing only the first few utterances (20%) of the help-seeker, suggesting it can be used to enhance early detection of suicide risk when deployed in the field. We analyze the approach with respect to different demographics, speaker focus and text truncation, and provide a few examples, qualitatively illustrating the benefits of our model. Our work goes the first step in helping counselors identify and treat at-risk individuals in real time.

Proceedings of the 29th International Conference on Computational Linguistics, pages 4241–4250
October 12–17, 2022.

Paper	*Embedding+Model*	*Language*	*Setting*
(Cheng et al., 2017)	LIWC+SVM	Chinese	social media (Weibo)
(Allen et al., 2019)	LIWC+CNN	English	social media (Reddit)
(Matero et al., 2019)	BERT without Pretraining	English	social media (Reddit)
(Ophir et al., 2020)	ELMO+Questionnaires+ANN	English	social media (Facebook)
(Lee et al., 2020)	W2V+LSTM+Lexicons	Korean	social media (Naver Cafe)
(Bantilan et al., 2021)	TF-IDF+XGBoost	English	phone counseling
(Xu et al., 2021)	Knowledge Graph+W2V+LSTM	Chinese	online counseling

Table 1: Sample of relevant approaches used for suicide risk classification from text.

2 Related Work

Our work extends past approaches to suicide risk detection in texts as well as NLP classification tasks in low-resource languages. For a systematic review of the use of machine learning for suicide risk detection from text we refer the reader to Ji et al. (2020) and to (Bernert et al., 2020). For a comprehensive survey of the application of the BERT architecture in different scenarios we refer the reader to (Rogers et al., 2020). In the remainder of this section we briefly survey recent works in each of these fields.

Detection of suicide risk There are limited works in suicide detection in conversations between help-seekers and counselors. Xu et al. (2021) used a classifier based on a knowledge graph of logical relationships of events related to suicide ideation. They combined this graph with Word2Vec embeddings to detect suicide risk in an online counseling service in Hong Kong. Their model achieved 81.5% ROC-AUC for suicide risk detection. Bantilan et al. (2021) combined TF-IDF embeddings with an XGBoost model in transcribed phone calls from a counseling service in English. Their approach achieved a 73% ROC-AUC performance in the phone call based counseling. None of these approaches addressed early detection, and are outperformed by our own approach in terms of ROC-AUC.

There is a body of work on suicide risk detection for English text from social media posts (Guntuku et al., 2017; De Choudhury et al., 2013; Zirikly et al., 2019). Online counseling chats are quite different from such settings in that they often include complete conversations between help-seekers and counselors, rather than single utterances, and exhibit temporal and mental-state dynamics.

Ophir et al. (2020), used ANNs to predict at-risk individuals from Facebook posts and psychological questionnaires. Matero et al. (2019) achieved top results in a suicide ideation detection task in social media (Zirikly et al., 2019) by adapting a BERT model to process input from Twitter and Reddit posts. They found that at-risk individuals use a distinct vocabulary in comparison to the rest of the population. Lee et al. (2020) tackled suicide ideation detection in social media posts in Korean which they describe as a low resource language. In their work they claim domain lexicons are highly beneficial for the task when available. Their classifying model is based on word embeddings, lexicons, attention, and LSTM. These works inspired our approach to combine expert based lexicons with the language model.

For convenience, the main approaches and the settings on which they were developed are summarized in Table 1.

Using Transformers for low-resource and morphologically rich languages Deep neural architectures, especially the Transformer, require massive corpora for adequate training. These pre-trained models are then fine-tuned for specific classification tasks, e.g., (Devlin et al., 2018; Sun et al., 2019; Pierse and Lu, 2020; Gururangan et al., 2020). However, fine tuning is suboptimal in the cases where the domain of the classification task is unique, especially in low resource and morphologically-rich languages (Klein and Tsarfaty, 2020; Seker et al., 2021; Nzeyimana and Rubungo, 2022). Our approach tackles those issues.

Hebrew NLP There is an increasing number of Hebrew tools available for modeling NLP tasks. Shapira et al. (2021) released a Hebrew Psychological Lexicons (HPL) that contains 30% of the terms that exist in LIWC, while also containing unique psychological terms that can help detect psychological aspects such as emotional state. Other tools include generic Hebrew BERT models such as HeBERT (Chriqui and Yahav, 2021) and AlephBERT (Seker et al., 2021). AlephBERT was trained on a larger dataset and achieved better results than HeBERT, making it our PLM of choice.

3 The Dataset

Sahar (https://sahar.org.il), Hebrew acronym for "Aid and Attention Online", is the leading chatline in Israel, focusing on suicide prevention, and emotional distress relief. Relieving the emotional distress of help-seekers is a crucial step in the process of suicide prevention (Overholser et al., 1997; Surís et al., 1996). The organization handles more than $10,000$ chat sessions a year, and these numbers have increased significantly during the COVID-19 pandemic (Zalsman et al., 2021).

Sahar's counselors are volunteers over 24 years old that completed a special training program by licensed clinicians. They are trained to use a special support language that is based on the conversation dynamics. During shifts, there are also therapists on duty who monitor the conversations and provide professional support if needed to the counselors. At the end of each session, the counselor is asked to summarize the conversation and to complete a short survey. A conversation is defined as exhibiting SI if the counselor answered "Yes" to the question "Did the subject of suicide come up in the conversation?"

The Sahar corpus The Sahar corpus contains 44,506 chat sessions which took place in the span of five years (2017-January 2022). Of these, 17,564 are labeled with a True/False to designate whether the chat exhibited SI. Seventeen percent $(3097/17564)$ of the labeled sessions are flagged with a positive SI label. For the remainder of this paper, the term Sahar dataset will refer to the 17,564 labeled sessions in the corpus. A session includes the utterances generated by the help-seekers and the counselor, delimited by a separating token. Table 2 shows general statistics related to our dataset.

Dataset statistics	
Total Num. Sessions	$44,506$
Num. Labeled Sessions	$17,564$
SI positive label ratio	17%
Mean(Median) number of tokens in a chat	$617(566)$

Table 2: General statistics for Sahar Corpus

Beyond the SI label, counselors are requested to select the prominent topics in each conversation (one to three topics chosen from a predefined list).

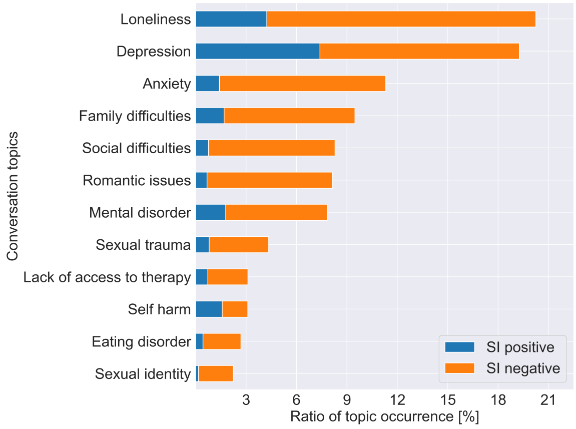

Figure 1: Conversation topics distribution in Sahar corpus

Figure 1 shows the discussed topics distribution in the data. As shown in the figure, loneliness and depression are the most common topics discussed in Sahar platform and a large share of those conversation exhibits positive SI. These findings coincide with psychology literature which depicts depression (Gijzen et al., 2021; Moitra et al., 2021) and loneliness (McClelland et al., 2020) as powerful predictors for suicide ideation. An interesting observation from Figure 1 is, that SI is prevalent across all of the reported topics, and is not exclusive to a certain topic.

Unfortunately, we are not able to share the Sahar corpus due to its sensitive and private content. Nor are we able to share the trained language model, since it was shown language models can be manipulated to reveal training data (e.g (Carlini et al., 2021)). We do provide however a repository with the experiments' code and lexicons used in this paper to support transparency and reproducibility.

4 Computational Approach

Our approach is based on an Ensemble method that extends a generic PLM to the SI domain, and leverages it with a fixed set of manually crafted (and clinically approved) set of suicidal cues. We expand on each component of the Ensemble model in turn.

4.1 The SI-BERT Classifier

SI-BERT is a model designed and configured for SI detection in online Hebrew chats. It utilizes AlephBERT (Seker et al., 2021), which is the best performing Hebrew BERT model to date and is publicly available. SI-BERT augments the generic AlephBERT model by (a) adding domain-specific to-

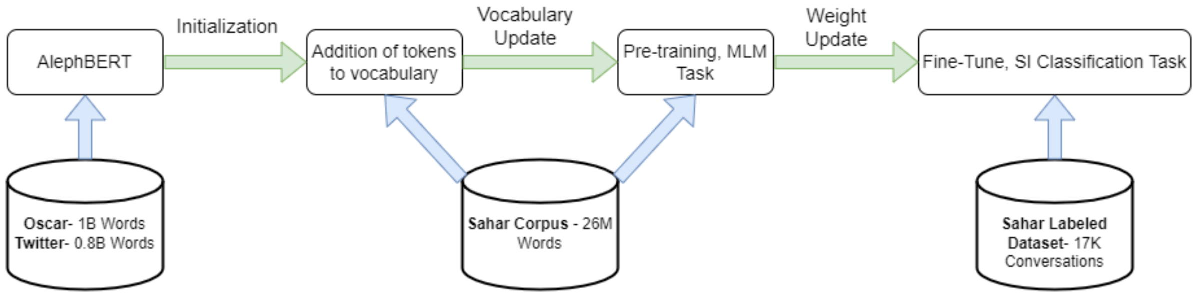

Figure 2: SI-BERT Architecture

kens to the vocabulary; (b) pretraining on a masked language model (MLM) task over the Sahar corpus; and (c) fine-tuning for the SI classification task. The SI-BERT architecture is illustrated in Figure 2. The term ‘SI-BERT’ at each step refers to the model obtained by the previous step. In the remainder of this section we expand on each of these steps.

Adding Domain Specific Tokens It is well established that enriching the vocabulary of pretrained models with Domain Specific Tokens (DST) improves performance for domain specific tasks (Tai et al., 2020; Beltagy et al., 2019; Honda et al., 2021).

We consider words which are OOV to be domain specific tokens, since by definition they exist in the domain corpus and not in the language model’s vocabulary. The language model’s vocabulary was constructed by taking the most common tokens from it’s training corpus. Manual examination of the most frequent DST finds that many of these tokens are highly related to suicide and mental distress e.g, “suicidal”, “desperate”, “depressed”, “abandoned” (translated from Hebrew). We hypothesize that adding a relatively small number of domain specific tokens will improve performance for suicide ideation classification task.

Therefore, the DST list we used is the δ most frequently appearing words in the Sahar corpus that were not in the vocabulary of the pretrained model. We added this list to SI-BERT’s vocabulary, changing it’s size from $|V|$ to $|V| + \delta$. We set $\delta = 1000$ based on performance on a held-out validation set. Analyzing the number of OOV token in Sahar corpus, we find 5% of the words to be OOV (before the addition of the DST list). After adding the DST to the vocabulary we observed a decrease of 20% in the number of OOV words in Sahar corpus.

Pretraining with MLM task In this step, we pretrain SI-BERT with a Masked Language Model (MLM) task. MLM is an unsupervised task in which a share of the tokens in each utterance is masked, and SI-BERT is trained to predict the masked words. This task has been shown to improve the performance of BERT and other language models performance for downstream tasks (Gururangan et al., 2020; Pierse and Lu, 2020). The training for the MLM task was conducted for 200 epochs on the complete Sahar corpus. An additional advantage of the MLM task is that it retrains the weights of the SI-BERT following the additions of tokens in the previous step (Tai et al., 2020).

Fine-Tuning In this step we fine-tune SI-BERT for the SI classification task. We add a binary classification head to SI-BERT. The classification head is a neural network layer which consists of two neurons with a softmax activation (binary classification). We compile the model with a cross entropy loss. We fine tune our model with the labeled dataset described in section 3. The BERT model is designed to process 512 tokens. We fine tune the model with the help-seeker’s text and used the first 512 tokens of each session as input to the model. For a discussion and justification of these decisions, see subsection 6.4. In practice, 21% of the sessions were truncated when inputted to the model.

4.2 Suicide Ideation Lexicon

We extracted 200 randomly selected positive SI sessions from the Sahar dataset and constructed a list of phrases that explicitly mention suicide ideation. The list contains 67 phrases such as: “suicide”, “cut wrists”, “want to die” and other variations.

This list was vetted by psychologists with expertise in suicide ideation. The set of 200 positive sessions was removed from the test sets used in the evaluation. Each session is mapped to a vector of length 67, where the ith element in the vector is the number of occurrences of phrase i in a given session. The vectors are scaled to [0,1] range and fed into a logistic regression model. We publicly share the lexicons in the article's repository.

4.3 Ensemble Model

The Ensemble model combines SI-BERT and the lexicon by feeding their predictions to a fully connected layer activated with a sigmoid function.

5 Evaluation

We compared the Ensemble model to the baseline models described below. The input to all models is a pre-processed chat that concatenates the utterances of the help-seeker and removes non-Hebrew characters and URLs. (See subsection 6.4 for a comparison with the case of also including the utterances of the counselor). For each bag of words based model (W2V, TF-IDF, HPL), the embeddings were scaled to [0,1] range and fed into a logistic regression model.

Fine-tuned AlephBERT (FT-BERT) We used the publicly available model of AlephBERT for text classification and fine tuned it on the labels in the Sahar dataset.

SI-BERT The SI-BERT PLM described in subsection 4.1.

Expert-based SI Lexicon (SI-Lexicon) The expert-based SI lexicon that is described in subsection 4.2.

W2V embeddings + Logistic Regression (W2V-LR) Word To Vector (W2V) is an algorithm which uses a neural network to map each word to a vector with a fixed size (Mikolov et al., 2013). We trained a W2V model on our corpus (embedding dimension=300, as used in the original paper (Mikolov et al., 2013)). The model was used to generate an embedding of all words in the session.

TF-IDF + Logistic Regression (TF-IDF-LR) Term Frequency–Inverse Document Frequency (TF-IDF) is a term weighting scheme commonly used to represent textual documents as vectors (Sammut and Webb, 2010). Each session is vectorized in this manner.

Hebrew Psychological Lexicon + Logistic Regression (HPL-LR) Each session is mapped to a vector of length 276 (number of lexicons in HPL). The ith element in the vector, is the number of occurrences of phrases in lexicon i in a given session.

We focus on two metrics to evaluate a model's performance: (a) ROC-AUC is the most commonly used metric in suicide detection tasks (Bernert et al., 2020). Its main advantage is that it doesn't depend on class distribution in the dataset. (b) The F2 metric computes a weighted harmonic average between the precision and recall scores. It assigns more than twice the weight (compared to the standard F1 metric) to the recall score which is sensitive to false-negative classifications. False negatives are critical in the SI detection task since missing people at risk has life-threatening consequences.

6 Results

In this section we report and discuss results from four perspectives: (i) Entire sessions results, (ii) Early detection, (iii) Results on different demographic groups, and (iv) Using the turn-taking structure vs. focusing on the utterances of the help-seeker. All results are reported using 5-fold cross validation. We keep the label imbalance unchanged in order to increase the potential use of the models for real time SI detection, discussing the False Positive-False Negative trade-off. The input to all models used in subsection 6.1 - subsection 6.3 consists of the concatenated utterances of the help-seeker in each session, while in subsection 6.4 we also provide results for a turn-taking scenario.

6.1 Entire Sessions Results

Table 3 compares the different models using several metrics, including ROC-AUC, and F2, which are commonly used for settings that suffer from high class imbalance (Forman and Scholz, 2010).

As shown in the table, the Ensemble model significantly outperforms the other models in terms of F1, F2 and ROC-AUC metrics. SI-BERT outperforms FT-BERT, and the other baselines. The SI-Lexicon achieves the highest precision, slightly better that the Ensemble, and does very well compared to the non-PLM approaches in the F1 and F2 metrics. However, it achieves only modest recall. This reflects the fact that the lexicon model was manually crafted by mental clinicians and tailored to detect explicit use of suicide ideation. However, there still exist SI positive sessions that cannot be captured by a static list of phrases.

The Ensemble model achieves a significant improvement in recall compared to the models it is

Model	*ROC-AUC[%]*	*F1[%]*	*F2[%]*	*Precision[%]*	*Recall[%]*
W2V-LR	86(0.25)	42(1.35)	33(1.50)	75(1.86)	29(1.50)
TF-IDF-LR	82(0.44)	43(0.30)	34(0.45)	75(1.50)	30(0.50)
HPL-LR	78(0.51)	28(0.94)	21(0.87)	66(2.53)	18(0.81)
SI-Lexicon	82(0.67)	51(0.58)	42(0.59)	**78(1.54)**	38(0.60)
FT-BERT	84(0.47)	47(1.35)	39(1.63)	70(1.61)	42(1.74)
SI-BERT	87(0.37)	55(1.21)	49(1.54)	71(1.52)	45(1.69)
Ensemble	**91(0.45)***	**61(0.89)***	**55(1.32)***	76(2.21)	**51(1.59)***

Table 3: SI classification results listing average performance with standard error in parenthesis. Bold highlights highest value, * marks a model is significantly better than the rest with $p < 0.05$ under Wilcoxon signed rank test.

comprised of, with only a slight decrease in precision performance. We wish to stress that this trade-off is of major significance in this specific domain. To further illustrate this point, Figure 3 presents false-negative ratios (out of test-set size) for the top performing models. As shown in the figure, the Ensemble model achieves the lowest false-negative (8.6%) given that the positive SI samples account for 17% of the test data. Most importantly, the Ensemble model reduces the false negatives ratio from 10.96% to 8.60% (11% decrease). In the field, such an improvement provides a meaningful contribution to suicide prevention, especially in early stages of the session, as we show in the following subsection.

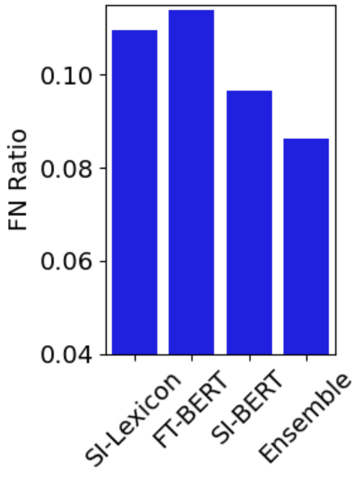

Figure 3: False-negative ratio (out of test-set size) for top-performing SI detection approaches.

6.2 Early Detection

Detecting at-risk individuals as early as possible during the session contributes to suicide prevention and reduces the load on the volunteers. To this end, Figure 4 shows the ROC-AUC performance of the top-performing SI-detection approaches when analyzing the first $\{5, 10, 20, 40, 60, 80, 100\}$ percent of the session (using 5-fold cross validation). As expected, all of the approaches improve as they process more information, with the Ensemble model constantly outperforming all of the other approaches.

Two key findings that stand out from Figure 4 are: (a) There is a consistent gap in performance between SI-BERT and FT-BERT, especially at early stages of the conversation. (b) SI-Lexicon performs poorly at an early stage of the conversation. We hypothesize that help seekers tend to be implicit, before allowing themselves to express their suicidal tendency explicitly[1]. Specifically, we found that while 73% of SI positive sessions contained an explicit SI phrase from the lexicon, only 38% of the SI positive sessions contained an explicit SI phrase in the early 20% of the session. This strengthens our conclusion that the lexicon model is insufficient for early detection of suicide risk.

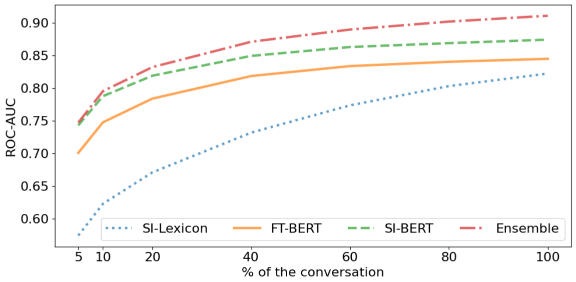

Figure 4: Classification results for early detection of top-performing SI detection approaches.

6.3 Demographic Analysis

Suicide risk, technological proficiency and linguistic norms vary across demographics. Therefore, we evaluate the performance of our model over different demographics.

[1] Furthermore, the explicit expressions could be a response to the counselor sensing implicit cues before directing the conversation a certain way.

Age/Gender	*Samples*	*+ Label*	*ROC-AUC*	*F2*	*Tokens*	*Types*	*OOV Tokens*	*OOV Types*
10-17	$4,179(23\%)$	15%	90	52	1.1M	54K	6%	6%
18-30	$9,066(52\%)$	19%	90	55	2.8M	141K	8%	7%
31-64	$4,164(24\%)$	17%	91	56	1.2M	107K	9%	6%
65+	$145(< 1\%)$	12%	93	58	29K	8K	12%	4%
Female	$12,074$	18	90	55	$3.6M$	$120K$	8%	8%
Male	$5,343$	17	91	55	$1.6M$	$83K$	8%	7%

Table 4: Ensemble model performance evaluation for subgroups of different age and gender.

	10-17	*18-30*	*31-64*	*65+*
10-17	–	0.28	0.39	0.86
18-30	0.60	–	0.55	0.91
31-64	0.54	0.37	–	0.89
65+	0.21	0.14	0.16	–

Table 5: Percentage of the number of unique tokens of each age group with respect to other age groups.

Age The Ensemble model consistently outperforms all other models across all age groups. Results of the Ensemble for different age groups are presented in Table 4 (Top), along with descriptive statistics, highlighting the differences between the sub-corpora in terms of size, types, tokens and OOV types and tokens.

The different linguistic norms each age group exhibits is captured in Table 5 through the relative size of each group's unique vocabulary. For example, while the sub-corpora 10-17 and 31-64 are a similar share of the data (23% & 24%, see Table 4) and a similar label break down (15% & 17%), 39% of the tokens used by the the 10-17 help-seekers are not used by help-seekers 31-64 years of age. Similarly, 54% of the tokens used by the 31-64 group are not used by the individuals on the 10-17 group. These trends are even more pronounced if one considers only OOV types.

Given the large variance in vocabularies between groups, the consistency of our results further demonstrates the robustness of our model.

Gender Examining the two main gender categories[2] we verify that the model achieves similar performance for both genders (see Table 4, Bottom). This result is far from trivial for two reasons: (a) Hebrew is a heavily gendered language (e.g., (Vainapel et al., 2015)). This means most (non-past tense) verbs and adjectives have different morphological inflection depending on the speaker's gender[3], and (b) The number of samples for each gender varies greatly, with female individuals making a vast majority of help seekers.

6.4 Speaker and Text Truncation

One major limitation of the BERT architecture is the constraint it enforces on length of the input. Consequently, most of the sessions cannot be processed fully. A straight forward way to tackle this constraint is to feed the model only part of the session – exhausting the 512-tokens buffer size. We considered three alternative protocols: (i) using utterances of both help-seekers and the counselor, (ii) using only utterances made by the help-seeker, and (iii) using only utterances made by the counselor. The latter protocol is used for comparative reason (also assuming that the responses of the counselor bear relevant signal). In each of these three settings, we considered two options (a) using the first 512 tokens ("keep head"), and (b) using the 512 trailing tokens ("keep tail").

Results of the Ensemble for each of the six settings are presented in Figure 5. The best performance is obtained using the head of the the utterances made by the help-seeker. This result, together with the ability to perform relatively well in early stages (subsection 6.2) are encouraging, given the high stakes of the task and the limited resources (personnel) available to the emergency services.

7 Discussion

The results in the previous section demonstrate the need for adequate adaptation of PLMs into a specific domain with obvious challenges in processing

[2] The gender the help-seeker identifies with is implicitly self disclosed since Hebrew is a gendered language and first-person verbs often takes different suffix according to the speaker gender (see examples in footnote 3).

[3] For example, consider the Hebrew inflections of the adjective 'lonely': 'boded' (M) vs. 'bodeda' (F), or the verb 'going (to)': 'holex' (M) vs. 'holexet' (F).

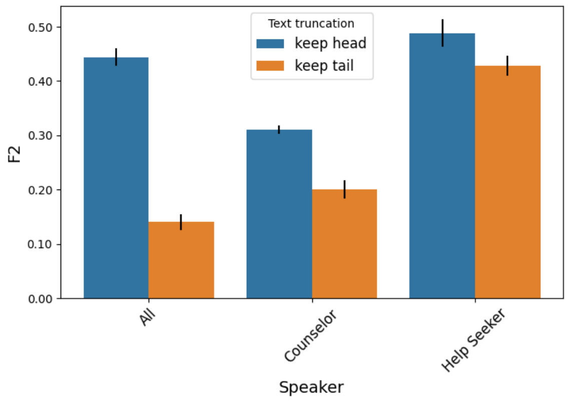

Figure 5: SI-BERT F2 performance for different text truncation methods and speaker text. Error bars mark standard error.

low-resource languages. Domain-tailored lexicons serve as strong baselines, with high precision and competitive F1 and F2 scores but they fall short in terms of recall and AUC. As PLMs tend to capture more nuanced expressions of SI – combining both approaches and careful fine-tuning improved the ability to detect SI early on in the chat.

We conclude this work discussing three illustrative examples (see Table 6) and reflecting on a few limitations of our approach.

I	*I don't want to die*
II	*I feel like life is too much for me*
III	*I had a spare time and bad thoughts, I decided to take a couple of sleeping pills and take a nap. I slept all day and now I'm dizzy.*

Table 6: Three illustrative utterances (translated from Hebrew). Note that these utterance are presented without a conversational context.

In utterance I (Table 6), the speaker explicitly rejects a suicidal intent, however, both the lexicon and SI-BERT (and the Ensemble, of course) classify this utterance as positive ideation. While the lexical approach matches the token *die* and wrongfully ignores the negation, our collaborators – clinical psychologists with suicide detection as their research focus – approve the classification, citing psychological studies on the distinction between suicide ideation and intent (Bagley, 1975; Beck et al., 1979; McAuliffe, 2002).

While it does not match any of the lexical items in the predefined lexicons, the second utterance (Table 6, II) is a classic example of SI. SI-BERT and the Ensemble (as well as the online counselor) correctly label it a positive SI, demonstrating the benefits of the domain-specific contextual model.

The third utterance is not considered by experts to exhibit SI. Indeed, SI-BERT (and the Ensemble) correctly classifies it as a negative example. On the other hand, FT-BERT assigns it a positive label. We hypothesize that the combination of "bad thoughts" and "(a couple of) sleeping pills" triggered the naive FT-BERT, while SI-BERT better captures the nuanced context. This example further demonstrates the benefits of fine-tuning the vanilla AlephBERT not only for the classification task but also fine-tuning the language model on the domain-specific data through a masked LM task.

We end this section briefly mentioning some limitations of the approach. First, our approach does not explicitly account for the discourse structure of the sessions. It may be that encoding the full conversational context may improve performance. Second, the Ensemble approach relies on a hand-crafted lexicon requiring considerable human effort. Many psychological lexicons already exist in other languages and have played a considerable role in prior SI research, see (Lee et al., 2020). Investing further effort in lexicon creation may have further reduced the false negative rate. Third, the lack of a benchmark dataset for the SI detection task makes it difficult to compare with prior work and approaches.

8 Conclusion and Future Work

Accurate and early detection of users' suicide risk in text-based counseling services is essential to ensure that at-risk individuals are given timely and proper treatment. This paper provides an automatic approach to risk detection from chats in Hebrew, a low-resource and morpholigically rich language. Our approach adapted a generic Hebrew language model by (i) adding out-of-vocabulary tokens, and (ii) performing additional pre-training of the LM on a masked language modeling task over the specific domain. Finally, we fine-tuned the model for the suicide risk detection task. We combined this model with a lexicon of hand crafted suicide ideation phrases that were vetted by experts. Our Ensemble model outperformed several competitive approaches, including a generic language model and the stand-alone lexicon. Our model performed consistently well for different demographics (age, gender). These encouraging results suggest the model can be deployed successfully, providing the much needed support to volunteers and health-care professionals in their mission of reducing suicide rates.

In future work, we wish to integrate the discursive structure into the model, and include latent information about the cognitive state of the help-seeker.

References

Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. Convsent at clpsych 2019 task a: Using post-level sentiment features for suicide risk prediction on reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 182–187.

Christopher Bagley. 1975. Suicidal behaviour and suicidal ideation in adolescents: A problem for counsellors in education. *British Journal of Guidance and Counselling*, 3(2):190–208.

Niels Bantilan, Matteo Malgaroli, Bonnie Ray, and Thomas D Hull. 2021. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. *Psychotherapy research*, 31(3):289–299.

Aaron T Beck, Maria Kovacs, and Arlene Weissman. 1979. Assessment of suicidal intention: the scale for suicide ideation. *Journal of consulting and clinical psychology*, 47(2):343.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Qijin Cheng, Tim MH Li, Chi-Leung Kwok, Tingshao Zhu, and Paul SF Yip. 2017. Assessing suicide risk and emotional distress in chinese social media: a text mining and machine learning study. *Journal of medical internet research*, 19(7):e243.

Avihay Chriqui and Inbal Yahav. 2021. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

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

George Forman and Martin Scholz. 2010. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *Acm Sigkdd Explorations Newsletter*, 12(1):49–57.

Mandy WM Gijzen, Sanne PA Rasing, Daan HM Creemers, Filip Smit, Rutger CME Engels, and Derek De Beurs. 2021. Suicide ideation as a symptom of adolescent depression. a network analysis. *Journal of Affective Disorders*, 278:68–77.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. 2021. Removing word-level spurious alignment between images and pseudo-captions in unsupervised image captioning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3692–3702, Online. Association for Computational Linguistics.

Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. 2014. Tracking suicide risk factors through twitter in the us. *Crisis*.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2020. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*.

Thomas Joiner, John Kalafat, John Draper, Heather Stokes, Marshall Knudson, Alan L Berman, and Richard McKeon. 2007. Establishing standards for the assessment of suicide risk among callers to the national suicide prevention lifeline. *Suicide and Life-Threatening Behavior*, 37(3):353–365.

Thomas E Joiner, Skip Simpson, Megan L Rogers, Ian H Stanley, and Igor I Galynker. 2018. Whether called acute suicidal affective disturbance or suicide crisis syndrome, a suicide-specific diagnosis would enhance clinical care, increase patient safety, and mitigate clinician liability. *Journal of Psychiatric Practice®*, 24(4):274–278.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2208–2217.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the sixth workshop on computational linguistics and clinical psychology*, pages 39–44.

Carmel M McAuliffe. 2002. Suicidal ideation as an articulation of intent: a focus for suicide prevention? *Archives of Suicide Research*, 6(4):325–338.

Heather McClelland, Jonathan J Evans, Rebecca Nowland, Eamonn Ferguson, and Rory C O'Connor. 2020. Loneliness as a predictor of suicidal ideation and behaviour: a systematic review and meta-analysis of prospective studies. *Journal of affective disorders*, 274:880–896.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Modhurima Moitra, Damian Santomauro, Louisa Degenhardt, Pamela Y Collins, Harvey Whiteford, Theo Vos, and Alize Ferrari. 2021. Estimating the risk of suicide associated with mental disorders: A systematic review and meta-regression analysis. *Journal of psychiatric research*, 137:242–249.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. *arXiv preprint arXiv:2203.08459*.

Yaakov Ophir, Refael Tikochinski, Christa SC Asterhan, Itay Sisso, and Roi Reichart. 2020. Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, 10(1):1–10.

James C Overholser, Stacy R Freiheit, and Julia M DiFilippo. 1997. Emotional distress and substance abuse as risk factors for suicide attempts. *The Canadian journal of psychiatry*, 42(4):402–408.

Nuo Wang Pierse and Jingwen Lu. 2020. Aligning the pretraining and finetuning objectives of language models. *arXiv preprint arXiv:2002.02000*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Claude Sammut and Geoffrey I Webb. 2010. Tf–idf. *Encyclopedia of machine learning*, pages 986–987.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Natalie Shapira, Dana Atzil-Slonim, Daniel Juravski, Moran Baruch, Dana Stolowicz-Melman, Adar Paz, Tal Alfi-Yogev, Roy Azoulay, Adi Singer, Maayan Revivo, et al. 2021. Hebrew psychological lexicons. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 55–69.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

Joan-Carles Surís, Nuria Parera, and Conxita Puig. 1996. Chronic illness and emotional distress in adolescence. *Journal of adolescent health*, 19(2):153–156.

Wen Tai, HT Kung, Xin Luna Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exbert: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439.

Sigal Vainapel, Opher Y Shamir, Yulie Tenenbaum, and Gadi Gilam. 2015. The dark side of gendered language: The masculine-generic form as a cause for self-report bias. *Psychological assessment*, 27(4):1513.

WHO. 2019. Suicide in the world: global health estimates. Technical report, World Health Organization.

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

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

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