Self-Adapted Utterance Selection for Suicidal Ideation Detection in Lifeline Conversations

Zhong-Ling Wang
Department of Computer Science
National Chengchi University
Taipei, Taiwan
109753106@g.nccu.edu.tw

Po-Hsien Huang and **Wen-Yau Hsu**
Department of Psychology
National Chengchi University
Taipei, Taiwan
psyphh@nccu.edu.tw hsu@nccu.edu.tw

Hen-Hsen Huang
Institute of Information Science
Academia Sinica
Taipei, Taiwan
hhhuang@iis.sinica.edu.tw

Abstract

This paper investigates a crucial aspect of mental health by exploring the detection of suicidal ideation in spoken phone conversations between callers and counselors at a suicide prevention hotline. These conversations can be lengthy, noisy, and cover a broad range of topics, making it challenging for NLP models to accurately identify the caller's suicidal ideation. To address these difficulties, we introduce a novel, self-adaptive approach that identifies the most critical utterances that the NLP model can more easily distinguish. The experiments use real-world Lifeline transcriptions, expertly labeled, and show that our approach outperforms the baseline models in overall performance with an F-score of 66.01%. In detecting the most dangerous cases, our approach achieves a significantly higher F-score of 65.94% compared to the baseline models, an improvement of 8.9%. The selected utterances can also provide valuable insights for suicide prevention research. Furthermore, our approach demonstrates its versatility by showing its effectiveness in sentiment analysis, making it a valuable tool for NLP applications beyond the healthcare domain.

1 Introduction

The suicide prevention hotline provides a free service for the people with mental issues to have the support on a 24/7 basis. On the other end of the call, a group of counselors who are trained volunteers are waiting for calls online in shifts. As mental health problems are becoming more and more prevalent in society, the difficulty of finding a sufficient number of counselors forms a common challenge that the suicide prevention hotline of many countries in the world is facing. The overwhelmed counselors and the busy phone line make the individuals in urgent need of help unable to get the opportunity to have the counseling. In order to alleviate of demand of suicide counseling, the intelligent conversation system forms a potential direction to study for aiding the hotline service. Fully automated chatbots as counselors providing counseling services are still impractical because the current natural language processing (NLP) technology has not yet been able to construct a reliable chatbot like a well-trained counselor to talk to these people who are mentally unstable and at high risk of suicide.

In this work, we approach the suicide counseling aiding from a more practical aspect. Instead of creating a chatbot to replace the human counselors, our goal is to propose a model for suicidal ideation detection (Ji et al., 2021). During the counseling, our model can be used for monitoring the conversation and identifying the suicide attempt of the caller. Once it is discovered that the caller is at a high risk of suicide or self-harm, more experienced counselors and other resources can immediately intervene at this time to prevent further harm.

Text-based suicidal ideation detection is a hot topic in the NLP area. In addition to medical records, clinical notes (Ji et al., 2021), and suicide notes (Schoene et al., 2021), previous work on textual suicidal ideation detection mainly focused on the social media data, such as tweets (Mishra et al., 2019; Sawhney et al., 2020; MacAvaney et al., 2021), microblogs (Huang et al., 2015), or online forums (Coppersmith et al., 2018; Sawhney et al., 2021; Yao et al., 2020; Alambo et al., 2019). Broadly monitoring the suicidality of users on the social media is no doubt a new way for public health professionals to prevent suicides. However, social media-based suicidal ideation detection can

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1436–1446
May 2-6, 2023 ©2023 Association for Computational Linguistics

cover only a part of cases as not everyone uses the social media and the users with suicidal ideation may not share their troubles publicly.

The scope of this work is entirely different from those based on social media data. We focus on the Lifeline conversations between callers in trouble and the well-trained counselors. Unlike the social media data, which are posted by the users unilaterally, the conversational data are made of the interaction between two or even more parties (e.g., the caller's family member). In the case of suicide prevention hotlines, a conversation is typically made by a caller and a counselor. The well-trained counselor will try to guide the caller to express her or his emotions and distress, revealing rich information for suicide risk assessment. The data used in this work were collected from the real world recordings provided by Taiwan Lifeline International.[1] Psychologists were invited for assessing and labeling the suicidal ideation of the caller in each conversation. To the best of our knowledge, this is a pioneering work of conversation-based suicidal ideation detection.

Unlike social media posts, the conversational data is usually longer and noisier. In our real world data, a conversation contains 515 utterances and 7,981 tokens in average, causing a barrier to models analyzing the content. In the long conversation, not all utterances provide the cue for suicide risk assessment. Instead of analyzing the whole conversation, we focus on key interactions between the caller and the counselor. However, extracting the important utterances from the hotline conversation can be still challenging due to the lack of sufficient training data. For this reason, we propose a novel self-adapted approach to utterance selection. Our idea is to focus on a number of discriminative utterances that provide useful information for NLP models to do the classification. To accomplish this goal, a distant-supervised model for suicidal ideation detection at the utterance level is further built for generating the training data for the utterance selection model.

Our approach to detecting suicidal ideation in phone-call conversations is distinct from reinforcement learning methods that are used to select important features or instances for training the main model. Instead, our distant-supervision methodology leverages information from the conversation level to the utterance level based on a key observation that the risk of suicidal ideation can often be determined by only a few utterances. Our approach also differs from previous hierarchical and multi-grained approaches (Kar et al., 2020; Angelidis and Lapata, 2018) that seek to distill fine-grained information from every smaller piece of the entire input.

In this work, we consider the suicidal ideation labels at the conversation level as the distant label and train a model to identify the discriminative utterances that lead to assessing the suicide risk of the caller. Due to the lack of annotation at the utterance level, we introduce a novel relabeling mechanism that is unique to our distant-supervision method. The result is a final model that can efficiently perform suicidal ideation detection on short, condensed inputs.

Our approach is evaluated on the real-world data labeled by psychologists, showing that the self-adapted utterance selection successfully improves the performance in suicidal ideation detection. We also explore our approach in another domain. Experimental results confirm that our approach is also effective in sentiment analysis, a typical NLP task. The contributions of this work are threefold:

1. We investigate the task of suicidal ideation detection in the phone-call conversations, introducing a new way to aid suicide prevention. Our system can assist the counselors in early detection of callers with high suicidal ideation, thus preventing suicides and reducing the work stress of the counselors.

2. We propose a novel self-adapted utterance selection approach that greatly condenses the conversations and successfully improves the model for dialogue understanding and other NLP tasks required to handle long input.

3. The utterances selected by our approach can also be taken as explanatory information that highlights key interactions between the caller and the counselor.

2 Related Work

This section summarizes the related work from three aspects, including the recent applications of NLP in suicide prevention, the conversation-based detection for other mental diseases.

[1] http://www.life1995.org.tw/

2.1 Suicidal Ideation Detection

The previous works on suicidal ideation detection were mainly built for the social media data due to the ease of data collection.

Gaur et al. (2019) investigated the detection of suicide risk based on a dataset consisting of 500 Reddit users. They based on the suicide risk severity lexicon to annotate the posts, and used Columbia Suicide Severity Rating Scale (C-SSRS) to determine the risk of suicidal ideation of each individual. Moreover, Mishra et al. (2019) extracted several kinds of features, such as Parts of Speech (POS), Latent Dirichlet Allocation (LDA), etc. from the tweets as the input of the trainer. Besides, linguistic Inquiry and word count (LIWC) is also a common method to analyze the utterance structure for individuals with mental problems. For instance, Shah et al. (2020) used LIWC for linguistic analysis on Reddit posts. Also, Schoene et al. (2021) performed LICW and presented the learning model to detect suicide notes on microblog, which figured out the syntax patterns of the individuals with mental problems. Despite these features provide part of the insight of utterances, most researchers combined these approaches to deep learning, which let studies more efficient and comprehensive.

In relation to deep learning for text-based data, bidirectional Long Short-Term Memory (Bi-LSTM) is also a common method for detecting mental health problem. Bi-LSTM is an approach that captures the frontend and backend features at every time step. Therefore, it provides much information from the utterances to make a detection. The study of ordinal suicidal ideation detection (Sawhney et al., 2021) designed an architecture, including Bi-LSTM layers, temporal attention layer, ordinal regression layer, to comprehensively analyze the posts from past. Then, they adapted C-SSRS to make an assessment of suicidality. Coppersmith et al. (2018) extracted contextual information between words via a Bi-LSTM layer, which handled with semantic features in the context.

Overall, broadly monitoring the suicidality of users on the social media is undoubtedly a new way to prevent suicide from public health. However, social media-based suicidal ideation detection can only cover a part of cases since not everyone uses the social media and the users with suicidal ideation may not share their troubles publicly.

2.2 Conversation-based of Mental Disease Detection

Dialogue analysis gains attention in the field of mental health care. Rinaldi et al. (2020) aim to predict depression based on the spoken data from interviews, and the multimodal approach (Zhang et al., 2021) that includes the audio features for depression detection in conversations. Other cognitive health issues can also be modeled in conversations (Farzana et al., 2020).

Clinical interviews are a widely utilized method for addressing mental health issues by professionals. However, the highly sensitive nature of much of this data makes it difficult for machine learning models to analyze. Despite these challenges, mental health-focused datasets are considered more consistent and reliable, leading to increased attention in the field of dialogue analysis for mental health care.

Xu et al. (2021) detects the suicidality from a 24-hour online text-based counseling services, which the form of dataset and the research objectives are similar with us. The researchers created the suicide knowledge graph to encode the utterance features as the input of Bi-LSTM layer, and then made a classification of the risk level with multilayer perceptron(MLP). In similar issue, Rinaldi et al. (2020) aim to predict depression in the interviews. They make an evaluation on the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014), which contains text transcripts and audio records of interviews for a clinical assessment for depression. In this study, they adopt the BERT (Devlin et al., 2019) model to classify depression or not on the conversation transcripts as the one of the baseline. However, their method is that each prompt of the interview is grouped by the probability distribution and concatenated with its belonging response, then converted to word embeddings using GloVe to transformed to word embedding as the features, and the result is obtained from the decision layer.

Apart from depression, dementia, such as Alzheimer's disease (AD), is also a hot topic of the dialogue based dataset. Farzana et al. (2020) adopt the Pitt corpus as the source of their dialogue act corpus. In addition to feature extraction, they build the classification models using support vector machine, decision tree, logistic regression, and neural networks. The result shows that the neural network model obtains the best score among those methods. Furthermore, Green et al. (2012) use the

Carolina Conversation Collection (CCC) corpus to facilitate conversation between the individuals with AD and their conversational partner. The CCC corpus includes recorded and transcribed conversations between researchers, students and individuals with AD. They aim to resolve the AD-related disfluencies via natural language understanding.

In these studies, the research focused on utterances from both or single side to figure out the context patterns of the mental health problem. Generally, utterances transformed by GloVe, BERT, etc. to word embeddings and the deep neural networks such as Bi-LSTM contribute to semantic analysis and are commonly used in nowadays. Though the results from the neural network models are difficult to observe the patterns of each mental disorder from the computing, we can combine with various features to make an analysis from both human and computing perspectives.

3 Dataset

We obtained the phone call recordings from Taiwan Lifeline International, and the audio recordings were transcribed by the experts with a psychology background. The transcriptions contain both counselors and callers' utterances, which are conversation-based forms. The transcriptions were further annotated by psychologists for assessing the caller's suicide risk level into five grades, including no suicidal ideation, occasional, frequent, ongoing suicide plan, and ongoing suicide attempt. The outlier conversations, such as not mention to suicide, cannot identify, statement not in person, and other situation will be removed. A high kappa value of 0.89 was measured, confirming a strong inter-rater agreement.

Due to the data sparsity, with experts' approval, we further merge the two weak ideation grades (i.e., no suicidal ideation and occasional) into low risk and the two strong ideation grades (i.e., ongoing suicide plan and ongoing suicide attempt) into high risk, resulting in three suicide risk levels. Table 1 summarizes the distribution of instances across different risk levels and presents the conversation length for each risk level in terms of the average number of utterances and tokens.

Risk Level	Overall	Low	Med.	High
# Instances	656	263	227	166
# Utterances	515	426	546	613
# Tokens	7,981	6,965	8,473	8,919

Table 1: Statistics of our dataset

4 Methodology

Given a conversation $\mathbf{x} = (x_1, x_2, ..., x_n)$, where x_i is either the prompt made by the counselor or the response made by the caller repeatedly, an intuitive way to predict the level of the caller's suicidal ideation can be defined in Equation 1.

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in Y} P(\mathbf{y}|\mathbf{x}) \tag{1}$$

where $Y = \{\text{Low Risk, Medium Risk, High Risk}\}$, and $P(\cdot)$ is a model that is trained to estimate the probability of y given $\mathbf{x}$.

As mentioned in Section 1, $\mathbf{x}$ can be very long and noisy, resulting in a poorer performance. Thus, our idea is to predict $\mathbf{y}$ given a short and condensed version of $\mathbf{x}$. That is, we aim to approximate $P(\mathbf{y}|\mathbf{x})$ by $P_C(\mathbf{y}|\hat{\mathbf{x}})$, where $\hat{\mathbf{x}} \subseteq \mathbf{x}$ and $P_C(\cdot)$ makes the prediction based on the subset of $\mathbf{x}$.

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in Y} P(\mathbf{y}|\mathbf{x}) \approx \arg\max_{\mathbf{y} \in Y} P_C(\mathbf{y}|\hat{\mathbf{x}}) \tag{2}$$

In Section 4.1, we present a novel self-adapted utterance selection approach that finds the utterances providing useful information that is easier for the underlying backbone model to capture. Section 4.2 shows the distant-supervised learning employed to train our model for utterance selection. Section 4.3 describes the combination of our framework, and Section 4.4 presents our model in the inference stage. Figure 1(a) and Figure 1(b) illustrate the details of our model in the training and the inference stages, respectively. Figure 1(c) shows the dataflow of our framework in evaluation.

4.1 Self-Adapted Utterance Selection

To extract the sentences that may contain essential suicidal ideation patterns from the conversation, we establish a model $P_S(\cdot)$ for finding $\hat{\mathbf{x}}$. The general form is given in Equation 3.

$$\hat{\mathbf{x}} = \{x_i | x_i \in \mathbf{x} \text{ and } P_S(x_i|\mathbf{x}) > \tau\} \tag{3}$$

where $P_S(\cdot)$ determines the importance of x_i to the whole conversation $\mathbf{x}$. In our case, however, $P_S(\cdot)$ cannot be directly trained in the supervised manner due to the lack of ground-truth. Thus, a novel distant supervised approach is proposed to estimate the importance of each utterance in a conversation.

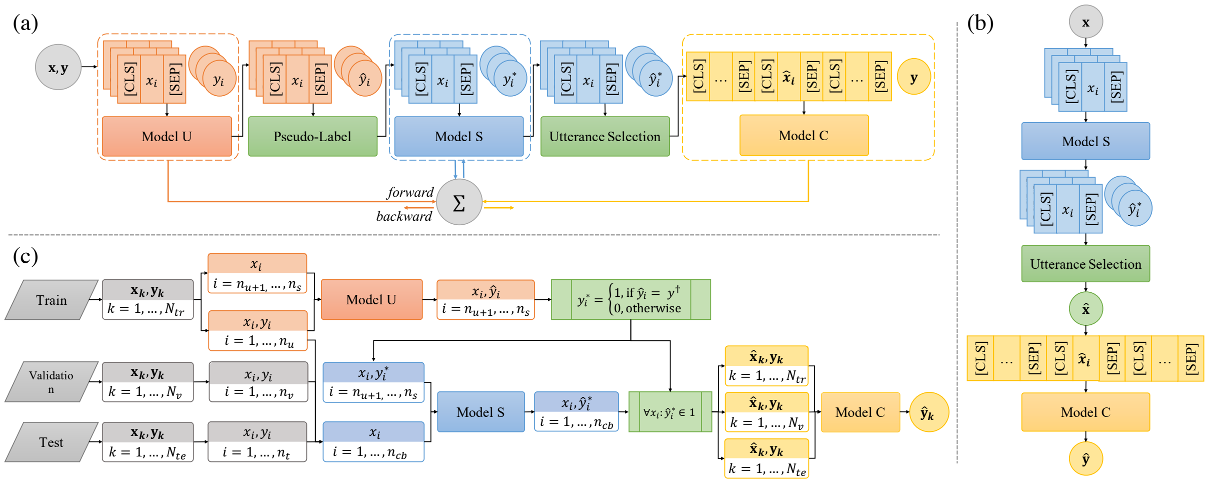

Figure 1: Overview of our approach. The sub-figure (a) shows our model in the training stage, the sub-figure (b) shows our model in the inference stage, and the sub-figure (c) shows the dataflow in evaluation.

Our assumption is that the important utterances are those able to contribute discriminative information for $P_C(\cdot)$ to predict the suicidal ideation. Thus, we aim to extract the utterances indicating suicidal ideation that is easily captured by NLP models. For this reason, we train another model $P_U(\cdot)$ that aims to predict the suicidal ideation at the utterance level.

$$\hat{y}_i = \arg\max_{y_i \in Y} P_U(y_i | x_i, \mathbf{x}) \tag{4}$$

where $\hat{y}_i$ is the suicidal ideation of the i-th utterance in $\mathbf{x}$. Note that $P_U(\cdot)$ is conditional on not only x_i but also its contextual information from $\mathbf{x}$.

For an utterance x_i from a conversation where the caller with a golden suicide risk level of $\mathbf{y}^\dagger$, we would like to pick up x_i as an important utterance if $\mathbf{y}^\dagger = \hat{y}_i$ because the underlying backbone model can make a consistent prediction based on the information in x_i. In other words, we prepared each training instance (x_i, y_i^*) for $P_S(\cdot)$ by determining the value of y_i^* as Equation 5.

$$y_i^* = \begin{cases} 1, & \text{if } P_U(y_i | x_i, \mathbf{x}) = \mathbf{y}^\dagger \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

The number of kinds of y_i^* is six for the binary condition at the three risk levels.

4.2 Suicidal Ideation at the Utterance Level

The last problem is how to train the model $P_U(\cdot)$. In our case, $P_U(\cdot)$ cannot be trained in the supervised manner because of the lack of labels at the utterance level. For this reason, we train it using distant supervised learning. For every utterance x_i in a conversation with a golden label of $\mathbf{y}^\dagger$ at the conversation level, we simply assign $y_i^\dagger$, the pseudo label of x_i, with $\mathbf{y}^\dagger$. For example, every utterance from a conversation where the caller was annotated with high suicide risk will be labeled with high suicide risk as well.

4.3 Training of the Whole Framework

The data preprocessing discards utterances that are longer than 29 tokens or shorter than 8 tokens, with an average token count of 15 tokens in reference to a single utterance. Role indicators, such as “A: ” for counselors and “B: ” for callers, are added to the beginning of each utterance. The models $P_S(\cdot)$ and $P_U(\cdot)$ are trained based on the pre-trained BERT model (BERT-BASE-CHINESE) (Devlin et al., 2019). On the other hand, $P_C(\cdot)$ is trained based on the pre-trained Longformer model (LONGFORMER_ZH) (Beltagy et al., 2020) for processing Chinese transcriptions, with an input length of 2,048 tokens. As shown in Figure 1(a) and Figure 1(b), the input of P_C is the concatenation of the selected utterances, while the input of P_S and P_U is a focused utterance and its neighboring utterances separated by `[SEP]`.

As illustrated in Figure 1(a), the three models $P_C(\cdot)$, $P_S(\cdot)$, and $P_U(\cdot)$ are updated in a sequential manner in each epoch. First, $P_U(\cdot)$ is updated to prepare the training data for $P_S(\cdot)$. Then, $P_S(\cdot)$ is updated to select utterances for training $P_C(\cdot)$. The losses of the models $\mathcal{L}_C$, $\mathcal{L}_S$, and $\mathcal{L}_U$ are calculated using cross-entropy, and the total loss $\mathcal{L}$ is their weighted sum which is used for updating the parameters of the three models.

4.4 Inference

Figure 1(b) shows our model in the inference stage. The $P_S(\cdot)$ and $P_C(\cdot)$, which obtain the lowest validation loss value from $P_C(\cdot)$, are saved for the inference stage. The $P_S(\cdot)$ model will shorten the original $\mathbf{x}$ to $\hat{\mathbf{x}}$, and the $P_C(\cdot)$ model predicts the level of the caller's suicidal ideation given $\hat{\mathbf{x}}$. Note that the $P_U(\cdot)$ model, which is built for training $P_S(\cdot)$, is not involved in the inference stage.

5 Experiments

We employ five-fold cross validation, with a development set split 10% from the training set. Significance testing is conducted using McNemar's test. The value of τ is set at 0.5, and the AdamW optimizer is utilized with a total of 40 epochs.

The baseline models include the vanilla BERT/Longformer models in the end-to-end supervised learning, the multi-turn GRU model, the JLPC model proposed by previous work for depression detection in interviews (Rinaldi et al., 2020), the LDA-based utterance selection method, reinforcement learning (RL) model, and the LIWC-based methods applying Gradient Boosting Decision Tree (GBDT) and multilayer perception (MLP) respectively, which are widely-applied in psychological tasks.

In the experiments, the input length with BERT is 512 tokens, and with the Longformer model is 2,048 tokens. The vanilla baseline is fed with the concatenation of all utterances formed as [CLS]$x_{counselor}$[SEP]x_{caller}[SEP]. The GRU model handles each turn in the conversation with the BERT/Longformer encoder separately and combines the information from the tokens with the GRU layer. JLPC model employs the latent prompt categorization on the utterances of counselors and also makes the prediction on first hundred pairs content. The LDA model performs the topic modeling for the conversation and clusters utterances into several groups. In brief, it is equivalent to let LDA model replace $P_U(\cdot)$ and $P_S(\cdot)$ models to do the utterance selection. We select the utterances from each cluster to form the input to the BERT/Longformer model and report the best performance by one of them.

Our novel distant supervised approach is also compared with reinforcement learning (Liu et al., 2019). We simplify our model to $P_S(\cdot)$ and $P_C(\cdot)$ as the framework of reinforcement learning. The $P_S(\cdot)$ as the agent is a binary classification to select the utterances according to the reward. The reward is defined as the difference from the loss of $P_C(\cdot)$ between two continuous epochs.

Different from the models based on Transformer, the LIWC-based model represent the conversation as a 79-dimensional bag-of-word vector, where the value of each dimension indicates the occurrences of the words belonging to the corresponding Chinese version of LIWC (Linguistic Inquiry and Word Count) category.[2] For each conversation, we perform word segmentation and obtain the distribution of LIWC categories. The MLP network is constructed for making the prediction as one of the baselines. GBDT classifiers which combine several weak learners and calculate with the gradient boosting decision tree algorithm are also adopted to accomplish the classification.

6 Results and Discussion

Table 2 shows the overall performances of our approach compared with baselines. The JLPC model making a binary classification for depression in previous work (Rinaldi et al., 2020) does not perform well on the hotline conversations. The GRU model with Longformer encoder produces the worst performance but has the better adaption with BERT encoder. The LIWC-based methods provide the information in different angle from these methods, achieving an F-score of 57.65% with GBDT classifier, higher than most of baselines, but the MLP network dose not perform well on the task. The LDA selection and RL methods figure out some important clues in conversations, which are superior to the vanilla BERT model. By contrast, the Longformer model does not benefit from these selection methods. The vanilla Longformer model can exploit the information from a large portion of the conversation and achieves an F-score of 59.77% in overall. However, our method with either BERT or Longformer significantly outperforms its respective vanilla model in terms of the overall performance at $p = 0.05$. Especially, our method with Longformer achieves the best F-score at overall and every risk levels, superior to all the other models. Particularly, our methodology accomplishes the highest gain on the high suicide risk, which is the most important case to be recognized and intervened. In the rest of this work, we adopt our method with Longformer as the subject for analysis in details.

[2] https://cliwc.weebly.com/

	Overall			Low Risk			Med. Risk			High Risk		
Method	P	R	F	P	R	F	P	R	F	P	R	F
BERT-based												
BERT	52.02	51.21	51.03	58.64	66.97	62.21	46.10	44.54	45.19	51.31	42.12	45.67
GRU	44.12	42.57	42.35	50.08	57.80	53.43	38.28	39.17	38.12	44.01	30.73	35.49
JLPC	40.45	42.41	39.57	43.79	49.99	44.17	39.62	41.13	39.85	37.93	36.11	34.70
LDA	54.59	52.77	52.45	58.44	63.58	59.99	50.53	47.21	48.16	54.81	47.52	49.21
RL	55.58	52.80	53.14	59.78	63.88	61.53	46.23	50.64	48.23	60.73	43.89	49.68
Longformer-based												
Longformer	60.38	59.93	59.77	69.02	64.19	66.01	55.30	57.68	56.24	56.81	57.91	57.04
GRU	41.03	38.49	36.30	44.51	65.69	52.51	36.97	32.89	33.58	41.62	16.88	22.81
JLPC	39.97	41.03	38.73	46.10	48.32	45.57	41.44	40.28	39.52	32.37	34.50	31.12
LDA	60.44	58.55	57.94	69.11	62.00	64.61	52.11	51.01	50.34	60.10	62.64	58.87
RL	59.21	56.67	56.75	65.18	63.50	63.34	51.29	52.93	51.88	61.15	53.58	55.03
Feature-based												
LIWC-MLP	50.26	49.26	49.48	54.69	51.70	52.82	41.78	44.90	43.16	54.29	51.18	52.47
LIWC-GBDT	58.93	57.28	57.65	63.42	71.47	67.14	49.77	49.83	49.67	63.61	50.53	56.15
Our Approach												
w/ BERT	59.29	58.74	58.53	68.71	62.34	65.19	53.96	57.28	55.12	55.20	56.60	55.28
w/ Longformer	66.58	66.50	**66.01**	73.01	69.54	**70.96**	62.74	60.77	**61.14**	64.00	69.20	**65.94**

Table 2: Performances of the three suicide risk levels and their macro-average, reported in Precision (P), Recall (R), and F-score (F) (%).

6.1 Choice of the Context Window

For an utterance $x_i \in \hat{\mathrm{x}}$ predicted as important by $P_S(\cdot)$, we can consider not only x_i itself but also its context, e.g., the previous and/or the next utterance of x_i. We concatenate x_{i-1} and/or x_{i+1} to x_i with the `[SEP]`. Table 3 compares the results of different context settings. The results indicate that incorporating additional context through utterances does not lead to improvement as the most relevant contextual information is already captured by the $P_S(\cdot)$ selection process.

6.2 Number of Selected Utterances

Table 4 shows how many utterances are extracted by our utterance selection approach compared with Longformer-based methods, including LDA, RL and original data. Based on the number of tokens in condensed conversations, the LDA and RL methods reduce the input length more than ours. Our utterance selection model $P_S(\cdot)$ reduces the average input length to 131 utterances and 2,405 tokens. Concerning the performance of vanilla Longformer baseline with data pre-processing, the model possibly lose the information, resulting in the worse result, an F-score of 55.23%. However, our method outperforms these approaches. Compared with other methods, our utterance selection model $P_S(\cdot)$ has more capability of identifying clues of suicidality from conversations.

6.3 Content of the Selected Utterances

Our approach allows us to understand what types of utterances are selected for the main model $P_C(\cdot)$ to predict the suicidal ideation. Through analyzing instances that were mispredicted by the baseline model but correctly predicted by our approach, we discovered that our $P_S(\cdot)$ extracts important patterns. Since Lifeline provides the service for people to confide their suffering, the negative words, such as “Bad mood”, “Painful”, “Wanna die”, are appeared in every risk level. In contrast, diverse daily life topics such as relationship, school, and work are mentioned at the low risk level. We can figure out what the main trouble occurred to the callers leads to the feeling of sadness or fear. Different from the low risk level, the thought of low self-esteem comes up in their words, such as “I am useless.” at medium risk level. Moreover, the strong hopeless, helpless and lonely feelings are revealed in the conversations. The word “death” is

	Overall			Low Risk			Med. Risk			High Risk		
Method	P	R	F	P	R	F	P	R	F	P	R	F
x_i	66.58	66.50	66.01	73.01	69.54	70.96	62.74	60.77	61.14	64.00	69.20	65.94
x_i, x_{i+1}	62.93	62.66	61.74	71.56	69.11	69.72	56.09	50.71	52.93	61.14	68.15	62.57
x_{i-1}, x_i	57.37	56.23	56.35	62.58	66.52	64.40	52.45	56.41	54.12	57.08	45.76	50.54
x_{i-1}, x_i, x_{i+1}	58.09	56.90	56.93	64.64	60.78	62.27	50.45	56.86	53.16	59.19	53.05	55.37

Table 3: With different context windows, the performances of our method at the three suicide risk levels and their macro-average, reported in Precision (P), Recall (R), and F-score (F) (%).

Selection Method	F-score	# Utter.	# Tokens
None	59.77	515	7,981
LDA	57.94	150	2,295
RL	56.75	218	3,930
Ours ($P_S(\cdot)$)	66.01	131	2,405

Table 4: Average lengths of the conversations condensed by different approaches. The macro-average F-scores (%) are reported.

usually in the utterances, but the suicidal ideation is only staying the thoughts.

Unlike above statements, physically suicidal means are mentioned at the high risk level. The callers usually talk about their suicide plan evidently, such as “I bought a helium barrel.”, “I took medicine and cut the wrist.”. The counselors then ask the caller’s location, e.g., “Where are you now?”, to ensure their safety. Overall, our approach effectively selects the utterances that are key to the interaction between the caller and the counselor, and provides valuable insights into the caller’s suicidal ideation.

6.4 Early Detection

Early risk detection plays an important role in mental issues (Wang et al., 2018). To reduce the burden for each counselor, we expect that the model has the ability to detect the risk of suicidal ideation as soon as possible so that the counselor can take a corresponding action to ensure caller’s safety. Figure 2 shows the F-scores of our method to do prediction given different numbers of first utterances in each conversation from the testing data. The black, blue, orange, and red dashed lines represent the F-scores of our method given all the utterances available as input in overall and at the low, the medium, and the high levels, respectively.

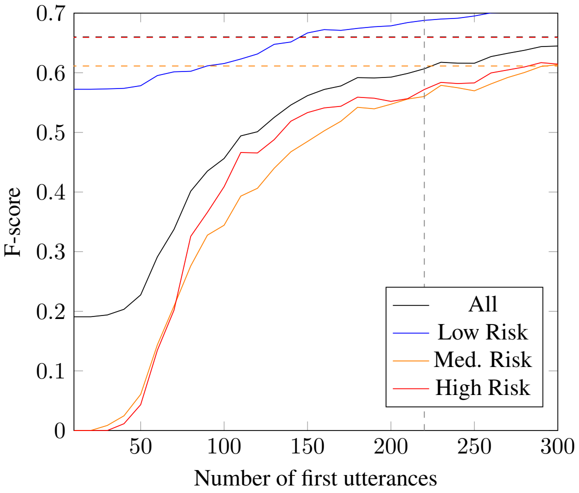

Figure 2: The figure shows the comparison of F-scores for our method in the scenario of early detection. The black, blue, orange, and red dashed lines represent the F-scores of our method with all utterances in overall and at low, medium and high risk levels, respectively.

Generally, the low risk cases are easier to identify. An F-score of 57.24% is achieved with only first 10 utterances being given to our method. The vertically dashed gray line represents the overall F-score achieves 60.67% in the first 220 utterances (i.e., 110 turns), superior to most of baselines. Based on our method framework, $P_S(\cdot)$ can directly detect each utterance whether it should be selected without any other information. The result of the early detection indicates that our method can identify the callers in different needs with a number of turns from the conversation, expected to distributing the suitable resources to assist them. Overall, our method not only contributes to extract the significant utterance patterns but also recognizes the important information during the conversations to accomplish early detection.

7 Applied to Sentiment Analysis

We also explore our approach in sentiment analysis, one of the most-studied application in NLP.

Method	F-score	# Tokens
BERT	86.62	718
Ours w/ BERT	89.63	266
Longformer	95.50	718
Ours w/ Longformer	92.74	371

Table 5: Experimental results of the Polarity dataset. The F-score (%) and the length of input fed into the backbone model are reported.

We evaluate our model for sentiment binary classification on the famous Polarity dataset (Pang and Lee, 2004), which consists of 1,000 positive and 1,000 negative movie reviews. Similar to our experiments, the split of data adopts five-fold cross validation. All methods are trained on the same split for five runs to verify the results.

Table 5 shows the F-scores of the baseline models and our methods. The average tokens fed into the backbone model is also shown. In respect to BERT-based models, our method achieves an F-score of 89.63%, superior to the vanilla BERT model obtaining 86.62%. However, though the performance of our method with Longformer model is improved, it is inferior to the vanilla Longformer model. The average length of the movie reviews in the Polarity dataset is 718 tokens, exceeding the input length of BERT but within the limitation of Longformer. Thus, the vanilla Longformer model is capable to handle full information from the input, while the BERT model can only exploit the information from the first 512 input tokens. Our method with BERT reduces the input length to 266 tokens in average, meeting the limitation of the BERT model and effectively selecting key sentences to enhance the performance.

These results support the key premise of our approach that by dramatically reducing the input length, our method can significantly improve the performance of the underlying model when the input exceeds its capacity. Furthermore, the results highlight the broader potential of our approach for other NLP tasks on lengthy and noisy data.

8 Conclusion

Suicidal ideation detection is one of the core interests in the domain of intelligent healthcare. This work explores the task on a new kind of data, the Lifeline conversations, and presents a novel self-adapted approach that condenses the long and noisy conversation for alleviating the model to identify the suicidal risk level of the caller. Experimental results on real-world data show our approach is not only significantly superior to existing models but also capable of applying to other domains. Our approach is also effective in the scenario of early detection, a practical issue for suicide prevention.

With the advancement of large language models such as ChatGPT, the development of intelligent mental health chatbots is becoming a reality in the near future. Our detection model can be integrated into these chatbots to trigger an emergency response if a high suicide risk level is detected.

Limitations

The Lifeline conversations are highly sensitive. Under the extreme circumstance where the callers suffer from suicidal ideation, it is inappropriate to ask them to consent to share their recordings publicly. For this reason, it is hardly to release a public dataset for the research community, resulting a barrier to reproducibility.

In this work, all the voice data were transcribed into spoken data by human. To launch a fully-automatic system, the transcription should be performed by machine, and more noises will be introduced by the speech recognition model. We will explore the voice-based approach to conversation-based suicidal ideation detection in the future.

Ethical Considerations

The data collected for this study is highly sensitive and contains personal information of the callers. To protect their privacy, all personal identifiable information such as names, ages, addresses, phone numbers, and places of work were removed during the transcription process. The data is securely stored and access to it is restricted to only authorized personnel. This study was reviewed and approved by the Human Research Institutional Review Board (IRB) at National Chengchi University[3] with the case number NCCU-REC-202102-006.

Acknowledgements

This work is partially supported by National Science and Technology Council, Taiwan, under grants 109-2222-E-001-004-MY3 and 110-2410-H-004-111-MY2. Thank you very much to the three anonymous reviewers and the meta-reviewer for your useful suggestions.

[3] http://rec.nccu.edu.tw

References

Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S Welton, and Jyotishman Pathak. 2019. Question answering for suicide risk assessment using reddit. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473. IEEE.

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota.

Shahla Farzana, Mina Valizadeh, and Natalie Parde. 2020. Modeling dialogue in conversational cognitive health screening interviews. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1167–1177, Marseille, France. European Language Resources Association.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The World Wide Web Conference*, WWW '19, page 514–525, New York, NY, USA. Association for Computing Machinery.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nancy Green, Curry Guinn, and Ronnie Smith. 2012. Assisting social conversation between persons with Alzheimer's disease and their conversational partners. In *Proceedings of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 37–46, Montréal, Canada.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in Chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562, Shanghai, China.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Sudipta Kar, Gustavo Aguilar, Mirella Lapata, and Thamar Solorio. 2020. Multi-view story characterization from movie plot synopses and reviews. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5629–5646, Online. Association for Computational Linguistics.

Miaofeng Liu, Yan Song, Hongbin Zou, and Tong Zhang. 2019. Reinforced training data selection for domain adaptation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1957–1968, Florence, Italy. Association for Computational Linguistics.

Sean MacAvaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80, Online.

Rohan Mishra, Pradyumn Prakhar Sinha, Ramit Sawhney, Debanjan Mahata, Puneet Mathur, and Rajiv Ratn Shah. 2019. SNAP-BATNET: Cascading author profiling and social network graphs for suicide ideation detection on social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 147–156, Minneapolis, Minnesota.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Alex Rinaldi, Jean Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18, Online.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7685–7697, Online.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2021. Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 22–30.

Annika Marie Schoene, Alexander Turner, Geeth Ranmal De Mel, and Nina Dethlefs. 2021. Hierarchical multiscale recurrent neural networks for detecting suicide notes. *IEEE Transactions on Affective Computing*, pages 1–1.

Faisal Muhammad Shah, Farsheed Haque, Ragib Un Nur, Shaeekh Al Jahan, and Zarar Mamud. 2020. A hybridized feature extraction approach to suicidal ideation detection from social media post. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 985–988.

Yu-Tseng Wang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2018. A neural network approach to early risk detection of depression and anorexia on social media text. In *CLEF (Working Notes)*.

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

Hannah Yao, Sina Rashidian, Xinyu Dong, Hongyi Duanmu, Richard N Rosenthal, and Fusheng Wang. 2020. Detection of suicidality among opioid users on reddit: Machine learning–based approach. *J Med Internet Res*, 22(11):e15293.

Pingyue Zhang, Mengyue Wu, Heinrich Dinkel, and Kai Yu. 2021. Depa: Self-supervised audio embedding for depression detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 135–143.