HKIE THE HONG KONG INSTITUTION OF ENGINEERS 香港工程師學會

Suicide risk level prediction and suicide trigger detection: a benchmark dataset

Jun Li[1,*], Xinhong Chen[2], Zehang Lin[1], Kaiqi Yang[1], Hong Va Leong[1], Nancy Xiaonan Yu[3] and Qing Li[1]

[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China
[2] Department of Computer Science, City University of Hong Kong, Hong Kong, People's Republic of China
[3] Department of Social and Behavioural Sciences, City University of Hong Kong, Hong Kong, People's Republic of China

ABSTRACT

The increasing number of suicide events in recent years has set off an alarm in society. To prevent suicides, it is important to develop platforms with associated intelligent algorithms, such as those for early suicide ideation detection (SID). In general, entering people's lives to obtain indicative information for effective SID is prohibitive due to the possible risk of privacy invasion. Social media posts provide valuable information about users' activities, indicating important hints toward SID in a non-intrusive manner. Although multiple datasets have been collected from social media platforms for efficient SID, they either neglect many suicidal posts by searching for pre-defined keywords, or contain limited information about suicide ideation. In this paper, a newly collected dataset, which expands the coverage of suicidal posts with more fine-grained annotations of suicide risk levels and suicide triggers compared with existing datasets, is presented. Benchmarking results by a popular deep-learning model are analysed to validate the reliability and potential of the collected dataset. Furthermore, multiple application scenarios of the dataset are discussed. The proposed dataset is expected to enhance the research on suicide ideation and behaviours, and have a strong impact on SID and more importantly, suicide prevention to save invaluable lives.

KEYWORDS Suicide ideation; suicide risk level prediction; suicide trigger detection; benchmark dataset; suicide prevention
CONTACT Xinhong Chen xinhong.chen@my.cityu.edu.hk
Received 28 March 2022

1. Introduction

Suicide ideation and behaviours are increasingly manifesting as a very serious societal problem, especially during the recent COVID-19 pandemic period. The World Health Organization has reported that an estimated amount of 700,000 people die by committing suicide every year. Many factors affect people's mental health, resulting in severe consequences such as mental health problems and suicide commitment (Ji et al., 2021). Among these factors, most can be alleviated or relieved via effective intervention programmes, and many suicidal tragedies can be avoided. Thus, it is necessary to develop effective tools for early detection and identification of potential suicidal ideation cases to allow early intervention. To this end, much progress has been made by researchers from two different communities. On the one hand, many systematic reviews and meta-analyses have been conducted by psychological experts to provide solid evidence supporting the findings of potential suicidal factors through statistical analysis (Amare et al., 2018; Ati et al., 2021; Castillo-Sánchez et al., 2020; Grimmond et al., 2019; Hassana, 1998; O'Connor and Kirtley, 2018; Werbart Törnblom et al., 2020). On the other hand, computer scientists have made promising efforts to design advanced machine-learning-based (ML-based) algorithms, the goal of which is to detect suicide ideation by considering different data sources, mostly social media posts containing an implicit or explicit expression of suicidal ideation (Gaur et al., 2019; Sawhney et al., 2018; Sinha et al., 2019). Despite the benefit brought by ML-based algorithms, existing works still suffer from the following two major drawbacks.

Firstly, ML-based algorithms demand a large amount of well-annotated data. However, the data collected by contemporary works for identifying suicide ideation is, quite often, done through keyword-based searching to build up the dataset, neglecting many potential suicidal posts and resulting in a limited and somewhat incomplete dataset. Besides, some popular datasets in the research community, such as the one constructed by Gaur et al. (2019), suffer from the utilisation of incomplete data coverage. In particular, only comments raised by other users are used to indicate the suicide ideation level of a poster. Rather, the original post inviting others' comments is disregarded. This incomplete and undesirable treatment can hardly reflect the actual ground truth. From the authors' viewpoint, a new dataset focusing on the users' original posts along with their hierarchical structure as reflected by the comments and responses to comments would be more desirable. Such a dataset would enable the building of a knowledge graph connecting the key entities and events with a rich structure, thereby facilitating feature selection and engineering, as well as the subsequent construction of effective machine-learning models.

Secondly, none of the existing works have paid attention to detecting the suicide triggers from social media posts. In reality, many suicide ideations only materialise in the presence of one or more suicide triggers. From the observations, it is less common for a user to post an explicit suicidal expression than some paragraphs containing

*The first author who was at the age of 35 or below on the closing date of submission to the HKIE Outstanding Paper Award for Young Engineers / Researchers 2022.

suicide triggers. Indeed, spotting possible suicide triggers can provide unique and essential information for effective suicide prevention, enabling an integrated study involving multiple vital factors pertinent to suicide to be conducted, well before the exhibition of suicide ideation and its subsequent detection in contemporary works.

To address the two drawbacks mentioned above, a cornerstone is the compilation of a new dataset from a public and popular platform – the Reddit website. Specifically, 139,455 posts from 76,186 users published under the "r/SuicideWatch" section from 01/2020 to 12/2021 have been crawled. Several pre-processing steps were then conducted to filter out noisy and irrelevant data since the focus was on suicidal ideation and behaviours. In order to obtain a comprehensive profile for each suicidal user, posts were collected from multiple Reddit sections, such as "r/depression", "r/selfharm", etc., pertaining to those users. As a result, the dataset finally contained 3,998 pertinent posts from 1,791 users.

The significance of the proposed dataset can be illustrated by multiple application scenarios, such as providing direct information for users' suicide risk levels, mining the users' suicidal factors from their daily posts, evaluating the effect of COVID-19 on the users' suicidal risk level change, etc. To facilitate such applications and produce ground-truth labels required for training ML-based models in a supervised manner, two types of labels were manually annotated for each of the posts collected in the dataset, namely suicide risk level and suicide triggers. Among the collected data, 500 random users were annotated in the first stage, leaving the remaining data to be published in the sequel. The agreement of the suicide risk level labels was measured by Fleiss' (1971) Kappa value, and the agreement of the suicide trigger labels was evaluated through a newly defined "hit rate" metric. Such manual annotation achieved an average Kappa value of 0.7782 and hit rate of 0.8609, which corroborates the soundness of the manual labels. With the ground-truth labels, the connection between the detected suicide triggers and the corresponding suicide risk level could be explored to inform their potential mapping patterns and further benefit each other. Based on the suicide risk level labels, analysis could be further conducted regarding the chronological trajectory of suicide risk level change, where event detection techniques (Li et al., 2017; Li et al., 2021) could be incorporated to study the causes of suicide risk level change. In a nutshell, this new dataset can play a vital role in achieving suicide prevention and provide a good foundation for further studies on suicide ideation detection and suicide prediction.

To demonstrate the utility of the newly proposed dataset, two application scenarios, namely suicide risk level prediction (SRLP) and suicide trigger detection (STD), are elaborated in this paper. Specifically, SRLP aims to correctly predict the suicide risk level based on a user's posts, and STD aims to detect suicide triggers from the post contents of each suicidal user. Based on a state-of-the-art language model, i.e., Bidirectional Encoder Representation from Transformers (BERT), a deep-learning model is implemented to address these two tasks. The BERT model is chosen here due to its wide adoption in obtaining meaningful representations for sentences with actual semantics. As the annotations of the suicide risk levels and suicide triggers naturally result in an imbalanced dataset (i.e., different categories contain significantly different numbers of cases or instances), extensive experiments were conducted on the proposed model under different training settings to resolve the imbalance issue. The experiment results demonstrate the validity of the annotation and the good utility of the collected dataset. In addition, the joint training of the two applications (i.e., SRLP and STD) promotes each other's performance, which testifies to the advantages of incorporating the suicide trigger annotation and detection task into the approach.

To sum up, the contributions of this work include the following aspects:

- A new dataset based on the Reddit platform is constructed and a structure with carefully annotated ground-truth labels is proposed; the dataset is well-suited to many valuable applications including SRLP and STD;
- Benchmarking results on top of the proposed deep-learning model are reported to demonstrate the utility and validity of the dataset annotations;
- Application scenarios and potential future directions based on the dataset are extensively discussed, thereby enhancing studies of suicide ideation and behaviour analysis for suicide prevention.

The remaining content of this paper is organised as follows. Section 2 briefly reviews existing datasets focusing on suicide ideation detection. Section 3 describes the details of the dataset collection and construction process, together with some valuable application scenarios on top of the dataset. Section 4 presents the proposed deep-learning model. Section 5 analyses the achieved benchmark results. In Section 6, several interesting and immediately available future applications based on this dataset are suggested and described. Finally, brief concluding remarks are offered in Section 7.

2. Related works

In recent years, many research studies have exploited social media platforms for the purpose of suicide ideation detection. These existing research works can be divided into Twitter/Facebook-based works and Reddit-based works according to the social platforms that they focus on.

For Twitter/Facebook-based works, most existing ones (Kour and Gupta, 2022; Rabani et al., 2020; Sawhney et al., 2018; Sawhney et al., 2021; Sinha et al., 2019;) first collect posts by searching for specific suicide-related keywords

(e.g., suicidal, depressed, etc.), and then manually filter out the suicide-related ones. Based on the remaining posts, users are further selected for annotation. For example, Vioulès et al. (2018) adopted a dictionary based on the risk factors defined by the American Psychiatric Association (Jacobs et al., 2003) and the warning signs identified by the American Association of Suicidology (Rudd et al., 2006). They defined four levels of labels (i.e., no distress, minimal distress, moderate distress and severe distress) to refine the different levels of suicide risk. However, these Twitter/Facebook-based datasets may neglect many implicit and important suicide-related posts as they are searched based on pre-defined lexicons.

As for the Reddit platform, existing works (Gaur et al., 2019,2021; Ji et al., 2018) collected data from the subreddit "r/SuicideWatch" and other suicide-related subreddits such as "r/selfharm" and "r/depression". Considering that some posts of "r/SuicideWatch" do not indicate any suicide risk, Aladağ et al. (2018) proposed to adopt a manual annotation strategy to filter out those posts with actual suicide risk. On top of Aladağ et al. (2018), Shing et al. (2018) further extended the categories into four levels (i.e., no risk, low risk, moderate risk and severe risk), but such a four-level-category design is not authoritative due to the lack of institutional endorsement as reported by Gaur et al., (2019). The Columbia Suicide Severity Rating Scale is then adopted to redefine five categories for annotation, namely supportive, suicide indicator, suicidal ideation, suicidal behaviour and actual attempt. Despite the adoption of these refined category labels, the dataset collected by Gaur et al., (2019) includes only users' comments on other users' posts instead of the users' own posts, containing rather limited suicide risk information about the users from the authors' viewpoint, a dataset containing each user's own posts should be collected and annotated to support suicide ideation detection more adequately.

3. Dataset construction

To address the limitations of existing datasets, more complete datasets from Reddit should be collected and annotation should be done more carefully and thoroughly. In this section, the dataset collection process, the manual annotation procedure and the two applications that were focused on are described in detail.

3.1. Dataset collection and pre-processing

Following the collection approach of the existing Reddit-based datasets, posts from the "r/SuicideWatch" subreddit section on the Reddit platform were first crawled. In particular, 139,455 posts from 76,186 users between 01/01/2020 and 31/12/2021 were collected. This specific time range was set with an emphasis on the period of the COVID-19 pandemic swamping the world. The impact on suicidal ideation and behaviour in this special period can provide unique information about how the virus affected individuals' mental health.

To facilitate the use of data in the subsequent annotation process, several pre-processing steps needed to be conducted. Firstly, to avoid privacy invasion, the collected data were anonymised by removing any users' identity-related data, such as names, addresses, emails and links. Then, overlapping posts or comments were removed from the same user. Considering that not all users under the "r/SuicideWatch" subreddit section exhibit suicidal intentions (Aladağ et al., 2018), the candidates of suicidal posts were determined by recognising the characteristic words in their posts. A combination of detecting negative expressions (Gkotsis et al., 2016) and identifying suicidally prominent terms (Sawhney et al., 2018) was conducted to filter out irrelevant users (i.e., those whose posts did not contain specific terms). An example of sentences containing negative expressions and suicidally prominent terms is "*I feel lonely and depressed all the time. I just want to end it all*".

On top of the filtered data, more information for each user was further collected since a user's suicide ideation tends to last for a period of time instead of just occurring at a single time point, and it can vary with time or interactions with other users. Procedurally, for each of the users, his/her last post was viewed as a representation of his/her latest *suicide ideation state*; hence, such a post is referred to as a "***targeted post***". Then, for each user's targeted post, the posts published before it and the comments under it were collected and summarised as *contextual information* of that targeted post. Note that the comments of other users were mostly encouraging the posters or echoing the posters' opinions, with little direct information concerning posters' suicidal ideation. As described in Section 2, the contemporary datasets based on these collected comments suffer from the coverage problem. Instead, these comments and the associated data have been utilised to build up the proper context of individual posts. They are valuable in mining the latent interaction between users in future work. In this paper, the focus is more on the post content due to space limitations. With more complete contextual information, a user's suicide ideation status within a period of time can be better detected and profiled, and a trajectory can be modelled. Besides, users' post not only under the subreddit "r/SuicideWatch" but also in many other subreddit sections. These relevant posts may also be informative. For example, the posts under "r/depression" may reflect the mental state information, which may not be explicitly mentioned under "r/SuicideWatch". Motivated by such an observation, the posts published within a specific time window were further crawled, e.g., a week preceding each user's targeted post from 14 relevant subreddit sections, such as "r/depression", "r/anxiety", "r/selfharm", "r/bipolar", "r/autism", "r/schizophrenia", "r/addiction", etc. Here, the choice of a week was based on preliminary

statistics, where the average gap between a user's latest post and the second-latest post is 3.63 days. Hence, a week's time can guarantee the coverage of multiple posts but avoid the inclusion of too many potentially noisy posts. Note that this period is flexible and can be adjusted as desired to create multiple versions of the dataset. In this study, one week was adopted as the period to conduct the preliminary research work. Finally, the current enriched dataset contained a total of 3,998 posts from 1,791 users.

Based on the enriched dataset, multiple application scenarios can be studied. For example, as the users' posts were collected instead of their comments, their suicidal risk levels can be directly and clearly detected through their post content. Mining the latent indicators of different suicide risk levels can help better predict users' mental health status and provide corresponding professional support. In addition, as users can freely post anything on social media platforms, it is more likely that the origins of their suicidal thoughts will be mentioned often in their historical posts. By extracting the triggers from the post contents and linking them to the users' suicidal risk levels, a more accurate prediction of users' suicidal risk levels can be made. Furthermore, the two-year period can be divided into more fine-grained epochs (e.g., a month or a quarter), and the users' suicide risk levels in these epochs can be evaluated, based on which the changes in the levels can also be studied to inform the causes of the changes. Due to space limitations, among those potential applications, the focus in this paper is placed on two of these applications (ref. Section 3.3) and the others are left to future work.

3.2. Dataset annotation

To benefit the multiple applications mentioned above, manual annotation was conducted on the collected dataset for the labels of suicide risk levels and suicide triggers. As each user has a targeted post and multiple historic posts, for each user, a post-level annotation was conducted based on the targeted post only, and a context-aware annotation based on all the posts published by the user. Due to the complexity and associated cost of the annotation work, in the first version of the published dataset, 500 random users were included for annotation, while the rest will be further annotated and published in the next edition.

Table 1. Detailed definition of different suicide risk levels.

Category	Notation	Definition
Indicator	IN	The post content has no explicit expression concerning suicide.
Ideation	ID	The post content has explicit suicidal expression but there is no plan to commit suicide.
Behaviour	BR	The post content has explicit suicidal expression and a plan to commit suicide or self-harming behaviours.
Attempt	AT	The post content has explicit expressions concerning historic suicide attempts.

The annotation scheme of the suicide risk level label is illustrated in Table 1, where the adopted labels and their corresponding annotation criteria have been approved by domain experts working in psychology (Gaur et al., 2019). Note that compared with Gaur et al. (2019), the "Supportive" category is deprecated as it was designed and used for the comment-based dataset to indicate that a user shows a positive attitude in his/her comments. The hierarchy of the adopted suicide risk level also echoes several existing suicide theories that suicide ideation goes from negative experiences or lacking a sense of belongingness to the accessibility of suicide, e.g., a suicide plan, tools, etc. (Lau et al., 2004; O'Connor and Kirtley, 2018; Wenzel and Beck, 2008).

As for building up the suicide trigger categories, theoretical studies on suicide (O'Connor and Kirtley, 2018; Wenzel and Beck, 2008) as well as recently published psychological systematic reviews and meta-analyses (Amare et al., 2018; Ati et al., 2021; Cheek et al., 2020; Gee et al., 2020; Grimmond et al., 2019; Hernández-Bello et al., 2020; Hill et al., 2020; Kahil et al., 2021; Ong et al., 2021; Surace et al., 2021; Werbart Törnblom et al., 2020) were referred to. After the integration of all the findings and an in-depth discussion with the sixth author (Nancy Xiaonan Yu), who is an expert in psychology and suicidal behaviour research study, 16 meaningful categories were included, i.e., *mental illness, physical illness, substance use, hopelessness, anxiety, low self-esteem, poor school performance, low socio-economic status, interpersonal violence, loss of loved ones, poor social support, interpersonal difficulty, dysfunctional family, closed ones' historic illness/suicide, unintended encounters (e.g., COVID-19 pandemic, car accidents, etc.)* and *sexual orientation-related issues*. The detailed coverage and the references of these categories are listed in Table A1 for better readability. In addition, an extra "Others" category (e.g., anti-capitalism, political factors) is included to make the classification complete. So far, there are $N_C = 17$ categories in total, including 16 trigger categories and the extra "Other" category. Unlike the suicide risk level annotation, sentence-level annotation was conducted for suicide triggers, by which more fine-grained information about the corresponding textual representation for each type of suicide trigger can be obtained.

To effectively conduct the above-mentioned annotation work, a group of three graduate students (i.e., the co-authors of this manuscript, Xinhong Chen, Zehang Lin and Kaiqi Yang) was recruited and trained by the first author (Jun Li) as annotators. They were divided into two groups, a group of Jun Li and Zehang Lin and another group of Xinhong Chen and Kaiqi Yang, each group labelling 250 users. During the annotation, the annotators were given the targeted and historic posts within a week preceding the targeted posts. Each annotator was required to give a post-level label and a context-aware label for the suicide risk levels, as well as the suicide trigger labels for the sentences of all the collected posts. Since the suicide

trigger annotation was done at the sentence level, the triggers annotated for the targeted posts only could be easily extracted and hence both post-level and context-aware suicide trigger labels could be established. Any disagreement within each group of annotators was resolved by an annotator from the other group (i.e., by the first and the second authors of this manuscript, Jun Li and Xinhong Chen). Due to space limitations, an example of the final dataset is given in Table A2, including the post content, the post-level annotations of suicide risk levels and suicide triggers and the context-aware annotations of the two labels.

To assess the quality of the annotations provided by the four annotators, different agreement metrics were adopted for the suicide risk level annotations and the suicide trigger annotations. For the suicide risk level annotations, Fleiss' Kappa value (Fleiss, 1971) was utilised. The overall average Kappa value of the post-level labels was **0.7866**, and that of context-aware labels was **0.7685**. Such high average Kappa values indicate the soundness and feasibility of the manually annotated suicide risk level labels. Statistics of the four annotated suicide risk labels are shown in Table 2. It can be observed that for both post-level and context-aware labels, the proportion of the AT category is the smallest, while the ones of ID and BR are the largest, which implies an imbalanced dataset matching the prevalence of suicidal ideation and behaviour well.

Table 2. Distribution of suicide risk level annotation.

Type	Post-level label				Context-aware label			
Distribution	IN	ID	BR	AT	IN	ID	BR	AT
	129	200	132	39	100	195	159	46

For the suicide trigger label agreement, the hit rate metric from marketing was adopted and adapted. Specifically, the hit rate is defined as $hr = \frac{N_{agree}}{N_{all}}$, where N_{agree} is the number of agreed suicide triggers labelled by an annotator and N_{all} is the number of labelled suicide triggers made by the annotator. Intuitively, a higher *hr* indicates that both annotators within the same group have more consistent annotation standards and reveals a higher quality. The average value of *hr* across all annotators was **0.8609**, which validates the reliability of the suicide trigger labels. The proportions of the 17 categories for the post-level and context-aware labels are shown in Figure 1. The figure shows that the following categories are more frequent: *mental illness, hopelessness, anxiety, low self-esteem, low socio-economic status, interpersonal violence, poor social support and dysfunctional family*. These results are consistent with previous findings of psychological systematic reviews and meta-analyses based on which the suicide trigger categories are built.

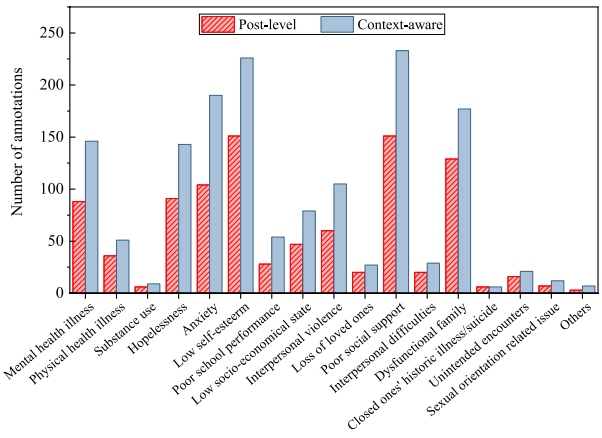

Figure 1. Category distribution of suicide trigger annotation.

More statistics concerning the suicide trigger annotation are shown in Table 3. It can be seen that most sentences did not contain a suicide trigger, which could make the detection of suicide trigger more difficult due to the sparseness of positive categorised annotations.

Table 3. Sentence-level statistics of suicide trigger annotation.

Type	Post-level	Context-aware
Total number of sentences	5,353	9,003
Sentences with triggers	799	1,241
Sentences with one trigger	666	1,028
Sentences with two triggers	104	168
Sentences with more than three triggers	29	45

3.3. Comparison with existing datasets

In the following, a systematic comparison is conducted between the newly constructed dataset and the existing ones. Four distinct differences are highlighted as shown in Table 4. First, the proposed dataset contains both the context-aware (i.e., "User" in the "Risk Level" column) and post-level annotations for suicide risk level (SRL) labels. Second, the dataset provides annotations of different granularity that can benefit applications with different goals. Third, the dataset is currently the only one containing fine-grained annotations of suicide triggers (ST), based on which the study of suicide cause and prevention can be greatly facilitated. As reported in Section 5.3, ST annotations can help further improve the prediction of suicide risk level. Fourth, the collection period of the dataset was set between 2020 and 2021, when COVID-19 spread everywhere; hence the collected data can be used to facilitate the study of the change in suicide causes and risk levels evoked by COVID-19. All the above-mentioned differences indicate that the proposed dataset is more desirable and advantageous to suicide-related studies.

Table 4. Dataset comparison.

Dataset	Source	Size	Annotation				Availability
			Risk level	Fine-grained SRL	Suicide trigger	COVID-19 covered	
UMD-RD (Shing et al., 2020)	Reddit	22,258 users	User	X	X	X	√
Ji (Ji et al., 2018)	Reddit, Twitter	5,920 posts	Post	X	X	X	X
Sina Weibo (Cao et al., 2019)	Sina Weibo	7,329 users	User	X	X	X	X
Twitter (Sinha et al., 2019)	Twitter	32,558 users	User	X	X	X	X
CLPsych (Zirikly et al., 2019)	Reddit	621 users	User	No/Low/ Moderate/ Severe Risk	X	X	√
SRAR (Gaur et al., 2019, 2021)	Reddit	500 users	User	Support / IN / ID / BR / AT	X	X	√
Ours	Reddit	500 users (3,998 posts)	Post, User	IN / ID / BR / AT	√	√	√

Note: (1) "User" in "Risk Level" column indicates context-aware labels; (2)"Availability" indicates whether the dataset can be accessed under the agreement of specific regulations; (3) "X" denotes that there is no such a setting in the datasets and "√" indicates otherwise.

3.4. Application scenarios

3.4.1. Suicide risk level prediction

As described above, the label set of suicide risk level and their distinct boundaries for these labels are shown in Table 1. For each user, a post-level and a context-aware label of suicide risk level can be allocated. Corresponding to these two labels, two important types of Suicide Risk Level Prediction (SRLP) task are considered: one solely based on the targeted post (referred to as "post-level SRLP") and the other based on all the collected posts (referred to as "context-aware SRLP").

Application 1 ***Post-level SRLP***: Given a user's targeted post, the goal is to predict a label indicating the suicide risk level of the user.

Application 2 ***Context-aware SRLP***: Given a user's targeted post and all the other posts preceding the targeted posts within a specific window (e.g., a week), the goal is to predict a label indicating the suicide risk level of the user.

Note that these two application scenarios differ in regard to the information need. With more complete contextual information, the user's recent suicidal risk level can be better recognised. However, in the absence of posts within the time window for collecting historic posts before the targeted post, i.e., when there are no other posts within the time window, the two tasks will coincide.

3.4.2. Suicide trigger detection

To further promote the early discovery of suicidal ideation and protection of suicidal people, another novel application is proposed, namely Suicide Trigger Detection (STD), which aims to detect the possible factors leading to suicide ideation from post content. Corresponding to the annotated labels, two relevant suicide trigger detection applications are proposed as follows.

Application 3 ***Post-level STD***: Given the targeted post of a user, the goal is to predict a set of labels for each sentence indicating the suicide triggers detected from that sentence.

Application 4 ***Context-aware STD***: Given a user's targeted post and all the other posts preceding the targeted posts within a specific window (e.g., a week in this paper), the goal is to predict a set of labels for each sentence indicating the detected suicide triggers.

4. Model design

Equipped with the collected dataset, two models are proposed for SRLP and STD respectively. The overall structure of the proposed model is shown in Figure 2. Considering that the two applications accept the same input but produce different output labels, *a word embedding*

module is shared for extracting the semantic features from the input clauses. In particular, the shared word-embedding module first operates on input sentences and converts the input words into latent vectors representing the linguistic semantics in an abstract form. Then, two machine-learning modules are designed to learn suitable classifiers from the dataset to produce the required output labels.

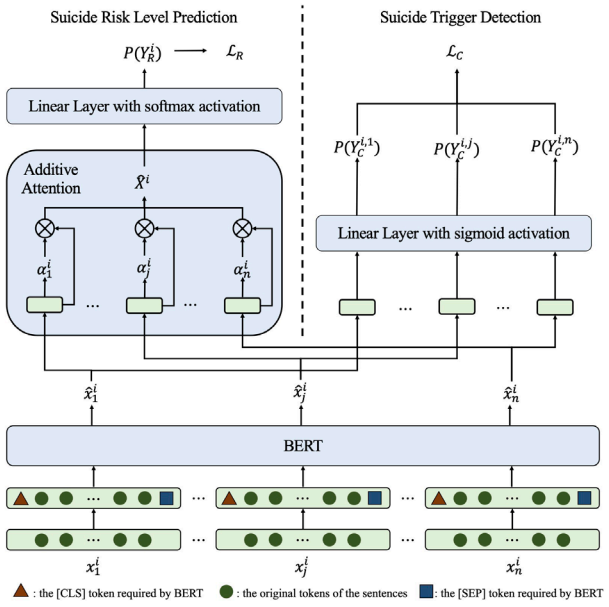

Figure 2. Structure of the shared model.

4.1. Word embedding module

The state-of-the-art language model BERT is adopted as the word embedding module. Specifically, given $X^i = \{x_1^i, \cdots, x_n^i\}$ and $x_j^i = \{x_{j,1}^i, \cdots, x_{j,\ell}^i\}$ with ℓ being the length of the sentence, the BERT model outputs $\hat{x}_{j,0}^i, \cdots, \hat{x}_{j,\ell+1}^i = \text{BERT}([CLS], x_{j,1}^i, \cdots, x_{j,\ell}^i, [SEP])$ for each sentence x_j^i and each sentence's embedding is empirically set to the first token's embedding $\hat{x}_j^i = \hat{x}_{j,0}^i$. Since BERT has been widely adopted in many contemporary works, interested readers are referred to its original paper (Devlin et al., 2019) for details.

4.2. Output layer and loss function for SRLP

As SRLP outputs a single label for each input post, an additive attention module (Bahdanau et al., 2016) is utilised on top of $[\hat{x}_1^i, \cdots, \hat{x}_j^i, \cdots, \hat{x}_j^i]$ to obtain a post level embedding vector, i.e., $\hat{X}^i = \sum_{j=1}^n \alpha_j^i \hat{x}_j^i$ and $\alpha_j^i = \frac{exp(W_a \cdot \hat{x}_j^i)}{\sum_{k=1}^n exp(W_a \cdot \hat{x}_k^i)}$, where $exp(\cdot)$ is the exponential function and W_a is a matrix of trainable parameters. The post level embedding vector $\hat{X}^i$ is then passed to a linear layer with *softmax* activation to predict the probability distribution of the suicide risk level $P(Y_R^i) = softmax\left(linear(\hat{X}^i)\right)$, where the *softmax* function can rescale the output value from the linear layer to a range of [0,1]. With $P(Y_R^i)$ computed, categorical cross entropy loss (Zhang and Sabuncu, 2018) is then adopted for updating of the model parameters, i.e., $\mathcal{L}_R = -\frac{1}{N}\sum_{i=1}^N \hat{Y}_R^i \cdot \log(P(Y_R^i))$, where N is the total number of posts used for training and $\hat{Y}_R^i$ is a ground-truth label vector. Note that this loss function is selected since each post can only be associated with a unique label. Then the above loss function is optimised through gradient descent by the AdamW optimiser (Loshchilov and Hutter, 2019).

4.3. Output layer and loss function for STD

Different from SRLP, for STD, each sentence context-encoded embedding vector $\hat{x}_j^i$ is passed to a linear layer with sigmoid activation to calculate the probability of each suicide trigger category, i.e., $P(Y_C^{i,j}) = sigmoid(\hat{x}_j^i) = [P(Y_C^{i,j,1} = 1), \cdots, P(Y_C^{i,j,N_C} = 1)]$, where *sigmoid* function is used since the respective probability of each suicide trigger should be calculated. Correspondingly, binary cross-entropy loss function (Zhang and Sabuncu, 2018) is adopted to calculate the prediction deviation between $P(Y_C^{i,j})$ and the ground truth suicide trigger labels $\hat{Y}_C^i$. Specifically, the loss function for STD is $\mathcal{L}_C = -\frac{1}{N}\sum_i \frac{1}{n}\sum_j \frac{1}{N_C}\sum_k [\hat{Y}_C^{i,j,k} \cdot P(Y_C^{i,j,k} = 1) + (1 - \hat{Y}_C^{i,j,k}) \cdot P(Y_C^{i,j,k} = 0)]$, and like SRLP, the loss function is optimised through the AdamW optimiser.

5. Experimental results

In this section, some experimental results on the annotated dataset for the two application scenarios are presented, and some insightful analysis of the current benchmark results is provided.

5.1. Evaluation metrics

Different evaluation metrics are utilised to demonstrate the feasibility and usefulness of the annotated dataset. For SRLP, since it can be viewed as a multi-class classification problem, F1 metrics are used for evaluation. More concretely, as the proportions of the four suicide risk levels are imbalanced as shown in Table 2, micro-F1 (*miF1*), macro-F1 (*maF1*) and the weighted-F1 (*wF1*) are adopted to measure the model performance as clarified by Opitz and Burst (2021). Among the three metrics, *miF1* is used to calculate the global classification performance, *maF1* is the

mean of the four individual F1 values of the four suicide risk levels, and *wF1* is used to take the proportion of the labels into consideration and return a weighted mean.

As for STD, the weighted F1 metric is also adopted for evaluation. Instead of the micro- and macro-F1 metrics, the mean Average Precision (mAP) metric is adopted, where the 11-point interpolated average precision (AP) (Craswell and Robertson, 2009) is used to calculate the average precision for each suicide trigger category, and mAP is the mean value of these APs.

5.2. Training and evaluation setting

As analysed above, the label distributions for both SRLP and STD are imbalanced, which could impact the machine-learning performance. To investigate whether or not balanced training can boost the prediction performance, the following three settings of reconstructing training data are explored and compared to simply training with the original data.

- Over-sampling (OSam) (Fujiwara et al., 2020): refers to the category with the largest number of instances, and duplicates the instances of the other categories to achieve a balanced training set
- Under-sampling (USam) (Fujiwara et al., 2020): refers to the category with the smallest number of instances, and randomly sample instances of the other categories to achieve a balanced training set
- Sample weighted loss function (SWL) (Ghosh and Lan, 2021): uses the inverse of the proportion of each category as its weight and constructs a weighted loss function. For example, the revised suicide risk level loss function is $\mathcal{L}'_R = -\frac{1}{N}\sum_{i=1}^{N}\frac{N}{N_{\hat{Y}_R^i}}\cdot\hat{Y}_R^i\cdot\log\left(P\left(Y_R^i\right)\right)$, where $N_{\hat{Y}_R^i}$ is the number of instances with the same label as the i-th one

To conduct a fair comparison, the following evaluation setting is adopted. The collected dataset is split into a training set and a testing set with a ratio of 8:2, and a generally adopted five-fold cross-validation in training deep learning approaches (Chen et al., 2020, Chen et al., 2021) is conducted to avoid the effect of randomness. Specifically, since the suicide risk level labels are imbalanced intrinsically, 20% from each category is sampled to form the testing set. As for STD, random splitting is adopted since the input of STD is a post instead of a sentence and it is impractical to sample at the sentence level. As for the five-fold cross-validation, the whole dataset is divided into five parts and the model is trained and tested five times, each time utilising one part of data for testing and the remaining data for training. Considering that the random sampling of USam training setting may introduce randomness, the experiment is repeated three times with different random seeds and the average performance is reported. In all the experiments, the number of training epochs is set to 50 and the learning rate is set to 10^{-5}, under which the model can smoothly converge to its best performance no matter which training manner it adopts.

5.3. Experiment results

The results of SRLP and STD are reported and analysed in the following. Table 5 depicts the performance of the proposed models on the annotated dataset under different training settings. As shown in the table, the proposed models with any training setting other than USam can achieve reasonable performance on SRLP. The consistently inferior performance under USam (with a standard deviation between 0.01 and 0.05) is due to the little training data. Such a result demonstrates the feasibility of the annotated labels well. Also, the current best performance for SRLP can only reach 0.5~0.6, which leaves ample space for improvement in future works and reveals the challenges of the proposed SRLP task. Interestingly, the proposed model appears to achieve a better performance only when using OSam. The reason behind this is that the adopted splitting naturally guarantees the competitive testing performance of an imbalanced dataset resulting from imbalanced training. Although intuitively the three data balancing settings should help the model avoid bias towards any category, the effect may not be revealed through the current training and testing split. Therefore, a further exploration concerning the necessity of balanced training would need to be performed. In addition, a preliminary trial of jointly training the two tasks was also conducted, as shown in Table 5 below with the row beginning with "SRLP+STD". Under this joint training, the two loss functions mentioned in Sections 4.2 and 4.3 are incorporated during training, i.e., $\mathcal{L} = \mathcal{L}_R + \mathcal{L}_C$. As can be observed from Table 5, the joint training of SRLP and STD results in higher weighted-F1 and micro-F1 values for all the training schemes, indicating that the inclusion of STD is very useful.

To examine the performance of SRLP in more detail, the classification performance of individual suicide risk levels is depicted in Table 6. The classification of the ID and BR categories is significantly better than the identification of the IN and AT categories. Such an observation echoes the category distribution reported in Section 3.2. In other words, the lack of sufficient training samples still hinders the effective learning of the proposed model for correct suicide risk level prediction. In future work, more data will be collected and annotated to upgrade the dataset so as to provide more information for prediction. In addition, by looking into the difference achieved in the two task settings, it is found that the performance is more balanced over the four SRL labels when training SRLP jointly with STD, i.e., the performance of IN and AT increases while that of ID and BR decreases a bit. This explains the increase of the weighted-F1 values and the decrease of the macro-F1 in Table 5.

Table 5. Performance of SRLP under different training settings on the annotated datasets.

Task	Training	Post-level			Context-aware		
		miF1	*maF1*	*wF1*	*miF1*	*maF1*	*wF1*
SRLP	Original	0.6199	0.6326	0.5640	0.5834	**0.6040**	0.5229
	OSam	0.6275	**0.6387**	0.5902	0.5521	0.5717	0.4903
	USam	0.4614	0.4622	0.4235	0.4328	0.4353	0.3935
	SWL	0.6118	0.6224	0.5657	0.5740	0.5898	0.5211
SRLP + STD	Original	0.6163	0.5920	0.6127	0.5838	0.5282	**0.5746**
	OSam	**0.6276**	0.5834	**0.6218**	0.5717	0.5097	0.5622
	USam	0.5061	0.5037	0.5001	0.4293	0.3908	0.4256
	SWL	0.6204	0.6042	0.6170	**0.5884**	0.5090	0.5672

Note: Bold underlined values represent the best performance.

Table 6. Detailed suicide risk classification performance under different training manners.

Task	Training	Post-level				Context-aware			
		IN-F1	ID-F1	BR-F1	AT-F1	IN-F1	ID-F1	BR-F1	AT-F1
SRLP	Original	0.4977	0.7017	**0.6762**	0.3805	0.3528	0.6715	**0.6777**	0.3897
	OSam	0.5063	**0.7071**	0.6592	0.4882	0.3327	0.6478	0.6325	0.3481
	USam	0.3516	0.5350	0.4934	0.3141	0.2449	0.5335	0.4540	0.3414
	SWL	0.4774	0.6964	0.6600	0.4288	0.3290	0.6601	0.6654	0.4301
SRLP + STD	Original	0.5018	0.6881	0.6172	0.5607	0.3593	0.6522	0.6523	**0.4487**
	OSam	**0.5163**	0.6931	0.6543	0.4698	**0.4323**	0.6388	0.6083	0.3593
	USam	0.4193	0.5510	0.4832	0.5611	0.2566	0.5194	0.4390	0.3479
	SWL	0.4907	0.6895	0.6294	**0.6073**	0.3786	**0.6916**	0.5869	0.3788

Note: Bold underlined values represent the best performance.

Table 7. Performance of STD under different training settings on the annotated datasets.

Task	Training	Post-level		Context-aware	
		miF1	*mAP*	*miF1*	*mAP*
STD	Original	0.4560	0.2214	0.4419	0.2064
	SWL	0.4510	0.2275	0.4384	0.2129
STD + SRLP	Original	0.6061	0.064	0.6663	0.0643
	SWL	0.6477	0.065	0.6509	0.0624

Note: Bold underlined values are the best performance.

Table 8. Category-wise AP values achieved by SWL training on the post-level STD.

Category	1	2	3	4	5	6	7	8	9
AP	0.4957	0.0866	0.0236	0.2476	0.4471	0.2792	0.3238	0.2235	0.5179
Category	10	11	12	13	14	15	16	17	
AP	0.2749	0.2994	0.3057	0.3070	0.0008	0.0320	0.0004	0.0011	

Note: The underlined values indicate that the trained model fails to detect these categories and hence, the task is really challenging even with the help of the state-of-the-art language model.

As for the performance of STD, the metrics values are reported in Table 7. The overall performance of STD is relatively low but still acceptable with 17 categories, considering the difficulty of the task. It can be observed that the joint training of STD and SRLP has greatly raised the micro F1 values of STD while the mAP metric is very low. Observations suggest that a more careful design is needed to handle the trade-off between the performance of SRLP and STD. For a clearer demonstration, the AP value of each category in the post-level STD application is reported in Table 8, which also matches the proportion of the suicide trigger distribution reported in Figure 1. The context-aware results are neglected due to similar observations. These results indicate that it is challenging to directly detect the suicide triggers from pure text content and that the lack of sufficient data samples may represent a tough barrier. Nevertheless, the task remains meaningful and beneficial. This is because if the suicide triggers can be detected early on, corresponding measures can be taken to avoid tragedy even well before detectable suicidal ideation comes to the surface. To sum up, further investigation into STD should be conducted to enhance its development.

6. Discussion

In this section, ethical consideration in view of the sensitive nature of suicide ideation detection is first discussed, and then some other potential application scenarios of the proposed dataset along with some future research directions are explored and suggested.

6.1. Ethical consideration

As the newly collected dataset comes from the Reddit platform, the users' privacy should be fully protected, even though Reddit is designed for anonymous posting. To this end, de-identification is performed on the collected data to remove any user-related information, such as names, addresses and emails (Sawhney et al., 2020). This results in a fully anonymised dataset, and the annotation conducted in this work has been kept separately from the raw user data to fully ensure user privacy and prevent misuse of data. All the example posts presented in this paper have been paraphrased, following the moderate disguise scheme proposed by Bruckman (2002). This research work is focused on predicting users' suicide risk levels and suicide triggers from the posted content in an observational and non-intrusive manner (Broer, 2020; Norval and Henderson, 2017). As such, there is no interference or diagnosis of users' experiences.

6.2. Future research directions

Based on the preliminary experimental results, some further issues as listed below will serve as potential future research directions:

- Prediction of suicide risk level: The experiments show that the adopted BERT-based neural network has already achieved a reasonable prediction performance. The performance is expected to be further enhanced if equipped with a more advanced model design (e.g., a sentence-level context encoder, etc.). Also, since the comments under every targeted post have also been collected in the dataset, a user-interaction graph based on the comment interactions between/among users can be constructed. The availability of such a graph will be beneficial to enhancing the feature learning for users, making use of the powerful *knowledge graph* paradigm.
- Suicide trigger detection and connection with suicide risk: The suicide trigger labels can be linked with the suicide risk levels and provide essential information for early interference and tragedy avoidance. Here, promising techniques can be proposed to address the imbalance and sparseness issues of suicide trigger in the social media texts. A possible solution can start by separating an STD task into two relatively easier sub-tasks, i.e., a binary classification task identifying whether or not a sentence describes a suicide trigger, and a multi-class classification task assigning a sentence with specific suicide trigger types.
- Suicide risk level trajectory generation and analysis: All collected posts belonging to the same user can be processed to form a suicide risk level trajectory for the user, where each footprint contains a suicide risk level. Especially during the COVID-19 period, acute stress levels and individuals' mental health may have been affected during quarantine lockdown, which may eventually have led to the formation of suicidal ideation or deterioration of mental health. Analysing the data from the temporal dimension and equipped with event detection techniques, the dynamics of the suicidal status over time can be modelled as a probabilistic finite state machine. Such a formulation can enable better study of the evolution of the suicidal levels of users over time.
- Suicidal causal relation detection: Based on the suicide trigger annotation, the detection range can be further extended to the causal relationships concerning suicidal ideation and behaviours. In view of the complexity of suicidal causality, an effective causal effect calculation technique can be proposed to evaluate how much each trigger contributes to an individual's suicide risk level. Based on the measured causal effects, a classification or regression task can then be proposed to facilitate the detection of the causal relationships and the inference of an individual's suicide risk.

7. Conclusion

In this paper, a new dataset that was crawled from the Reddit platform has been put forward. This new dataset is coupled with manual annotations so as to support important research tasks like suicide ideation detection and suicide prediction. Compared with existing datasets, the dataset in this study covers a more recent period (i.e., the COVID-19 period), contains a more direct source (i.e., Reddit users' posts instead of only the comments) and features more fine-grained annotations of both suicide risk levels and suicide triggers. To demonstrate the utility and validity of this new dataset and its annotation serving as the ground-truth labels, a deep learning model is proposed and several experiments are conducted, with remarkable benchmark results and agreement metrics. To further promote the studies on suicide ideation and behaviour analysis upon the proposed dataset, several application scenarios have been considered, and future research directions outlined. It is envisioned that with the help of this newly proposed dataset, research on advanced suicide trigger detection can be conducted, and effective and intelligent suicide prevention systems can be designed and implemented so as to prevent suicidal tragedies in a timelier manner.

Appendixes

Table A1. Specifications of suicide trigger categories.

Index	Category	Definition or Coverage	Reference
1	Mental health illness	Any existing mental disorder and mental illness, including but not limited to depression, personality disorder, eating disorder, PTSD, social phobia, etc.	Amiri, 2020; Carrasco-Barrios et al., 2020; Demenech et al., 2021; Kahil et al., 2021; Surace et al., 2021
2	Physical health illness	Any existing physical disorder, physical defects, including but not limited to cancer, reading disorder, COVID-19 symptoms, obesity/underweight, less sleep time, etc.	Amare et al., 2018; Amiri, 2020; Ati, Paraswati and Windarwati, 2021; Richardson et al., 2021
3	Substance use	The uncontrollable use of drugs/alcohol/tobacco.	Carrasco-Barrios et al., 2020; Grimmond et al., 2019; Hernández-Bello et al., 2020; Ong et al., 2021
4	Hopelessness	Hopelessness, pointlessness, meaninglessness, emptiness, feeling stuck/trapped/in a cycle, etc.	Amare et al., 2018; Gee et al., 2020; Méndez-Bustos et al., 2019; Miranda-Mendizabal et al., 2019
5	Anxiety	Anxiety, fear, worry, stress, etc.	Demenech et al., 2021; Howarth et al., 2020; Kahil et al., 2021; Pengpid and Peltzer, 2020; Werbart Tornblom et al., 2020
6	Low self-esteem	Feeling like a failure, worthless, ugly, a loser, a burden, etc.	Gee et al., 2020; Grimmond et al., 2019; Miranda-Mendizabal et al., 2019
7	Poor school performance	Low school performance, fail school tests	Amare et al., 2018; Grimmond et al., 2019; Hernández-Bello et al., 2020; Werbart Törnblom et al., 2020
8	Low socio-economic status	Unemployment, getting fired, having no money, kicked out of home, homeless, etc.	Carrasco-Barrios et al., 2020; Howarth et al., 2020; Jafari et al., 2020; Pengpid and Peltzer, 2020; Richardson et al., 2021
9	Interpersonal violence	School/workplace/internet/anywhere harassment/ bullying (verbal or physical)/abusive behaviour	Angelakis et al., 2020; Hernández-Bello et al., 2020; Ong et al., 2021; Surace et al., 2021; Werbart Törnblom et al., 2020
10	Loss of loved ones	Breakup, divorce, losing relatives or friends (by death)	Duarte et al., 2020; Howarth et al., 2020; Jafari et al., 2020; Richardson et al., 2021
11	Poor social support	Lose/lack friends, have no one, loneliness, isolation, being rejected, being neglected, being abandoned, etc.	Amare et al., 2018; Duarte et al., 2020; Gee et al., 2020, 2020; Kahil et al., 2021; Miranda-Mendizabal et al., 2019; Richardson et al., 2021; Werbart Törnblom et al., 2020
12	Interpersonal difficulties	Don't know how to make friends/socialise; having difficulties in socialisation	Ati et al., 2021; Miranda-Mendizabal et al., 2019; Richardson et al., 2021
13	Dysfunctional family	Family conflict, poor support from family members, parents' dereliction of duty, indifferent family, etc.	Ati et al., 2021; Hernández-Bello et al., 2020; Miranda-Mendizabal et al., 2019; Surace et al., 2021; Werbart Törnblom et al., 2020
14	Close ones' history of illness/suicide	The historic illness or suicide experiences of friends or families	Ati et al., 2021; Carrasco-Barrios et al., 2020; Hill et al., 2020; Ong et al., 2021
15	Unintended encounters	Pregnancy, robbery, uncertain future due to pandemic, etc.	Howarth et al., 2020; Méndez-Bustos et al., 2019; Miranda-Mendizabal et al., 2019;
16	Sexual orientation-related issues	Gender/sexual disorder, sexual orientation related suicidal thought, same-sex relationship	Richardson et al., 2021

Table A2. Examples of different suicide triggers and associated risk levels.

Type	Content	Suicide trigger	Suicide risk level
Targeted post	Life is worse than death.		Post-level: Ideation Context award: Attempt
	Right now I have so many bad memories and thoughts in my head and nobody listens to me and my dad just yells at me and gives off this feeling of being tired of me and my mother is always occupied with her boyfriend and doesn't give a flying fuck about me.	dysfunctional family	
	I have this hopelessness feeling like I'm stuck inside a nightmare but I don't wake up and ugh.	hopelessness	
	I'm sorry but right now I'm considering suicide because I can't fucking take it anymore.		
Additional post	Well I just started and can't stop.		
	Yeah well, some days ago I started cutting my arm and I seriously can't stop doing it, idk why tbh.		

Acknowledgements

This work was supported by the Hong Kong Research Grants Council under a Collaborative Research Fund (Project No. C1031-18G).

Notes on contributors

Mr Jun Li received his Master's degree from Sun Yat-sen University (SYSU), Guangdong, China in 2019. He is currently a Ph.D. candidate in the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. His research interests include weakly-supervised machine learning and robustness in machine learning.

Dr Xinhong Chen obtained his B.Eng. from Sun Yat-Sen University in 2018 and his Ph.D. degree from Department of Computer Science at City University of Hong Kong. His research interest includes causality analysis, natural language processing, sentiment analysis and autonomous driving.

Mr Zehang Lin obtained his Bachelor's and Master's degrees from Guangdong University of Technology (GDUT), Guangdong, China in 2019. He is currently a Ph.D. candidate in the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, People's Republic of China. His research interests include domain adaptation and cross modal retrieval.

Dr Hong Va Leung obtained his B.Sc. and M.Phil. degrees from The Chinese University of Hong Kong in 1987 and 1990 respectively, and received his Ph.D. degree from the University of California in 1994. He is currently an associate professor and the associate head of the Department of Computing at The Hong Kong Polytechnic University. His research interests lie in distributed systems, distributed databases, mobile computing, applications originating from the domains of parallel programming, internet and geographical information systems.

Dr Nancy Xiaonan Yu is a health psychologist with training in public health. Her research programme has focused on resilience in adversities by examining the resilience process and associated factors at the individual, dyadic, family and community levels. In addition, she has developed resilience-based intervention programmes to promote mental health in stressful populations.

Prof Qing Li received his B.Eng. degree in Computer Science from Hunan University, Hunan, China in 1982, and M.S. and Ph.D. degrees in Computer Science from the University of Southern California, L.A., California, USA in 1985 and 1988 respectively. He is currently the Chair Professor and the head of the Department of Computing at The Hong Kong Polytechnic University. His research focuses on data science, web mining and artificial intelligence.

References

[1] Aladağ AE, Muderrisoglu S, Akbas NB, Zahmacioglu O and Bingol HO (2018). Detecting Suicidal Ideation on Forums: Proof-of-Concept Study. *The Journal of Medical Internet Research*, 20(6), e215.

[2] Amare T, Woldeyhannes SM, and Yeneabat T (2018). Prevalence and Associated Factors of Suicide Ideation and Attempt among Adolescent High School Students in Dangila Town, Northwest Ethiopia. *Psychiatry Journal*, 2018, 7631453.

[3] Amiri S (2020). Prevalence of Suicide in Immigrants/ Refugees: A Systematic Review and Meta-Analysis. *Archives of Suicide Research*, 26(2), pp. 1–36.

[4] Angelakis I, Austin JL and Gooding P (2020). Childhood maltreatment and suicide attempts in prisoners: a systematic meta-analytic review. *Psychological Medicine*, 50(1), pp. 1–10.

[5] Jacobs DG, Baldessarini RJ, Conwell C, Fawcett JA, Horton L, Meltzer H, Pfeffer CR and Simon RI (2003). Practice guideline for the assessment and treatment of patients with suicidal behaviors. *The American Journal of Psychiatry*, 160(11 Suppl), pp. 1–60.

[6] Ati NAL, Paraswati MD and Windarwati HD (2021). What are the risk factors and protective factors of suicidal behavior in adolescents? A systematic review. *Journal of Child and Adolescent Psychiatric Nursing*, 34(1), pp. 7–18.

[7] Bahdanau D, Cho K and Bengio Y (2016). Neural Machine Translation by Jointly Learning to Align and Translate, *ArXiv Preprint*, arXiv:1409.0473.

[8] Broer T (2020). Technology for Our Future? Exploring the Duty to Report and Processes of Subjectification Relating to Digitalized Suicide Prevention, *Information*, 11(3), 170.

[9] Bruckman A (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology*, 4(3), pp. 217–231.

[10] Cao L, Zhang H, Feng L, Wei Z, Wang X, Li N and He X (2019). *Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention*. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, pp. 1718–1728.

[11] Carrasco-Barrios MT, Huertas P, Martín P, Martín C, Castillejos C, Petkari E and Moreno-Küstner B (2020). Determinants of Suicidality in the European General Population: A Systematic Review and Meta-Analysis. *International Journal of Environment Research and Public Health*, 17(11), 4115.

[12] Castillo-Sánchez G, Marques G, and Dorronzoro E (2020). Suicide Risk Assessment Using Machine Learning and Social Networks: a Scoping Review. *Journal of Medical Systems*, 44(12), 205.

[13] Cheek SM, Reiter-Lavery T and Goldston DB (2020). Social rejection, popularity, peer victimization, and self-injurious thoughts and behaviors among adolescents: A systematic review and meta-analysis. *Clinical Psychology Review*, 82, 101936.

[14] Chen H, Yang J and Chen X (2021). A convolution-based deep learning approach for estimating compressive strength of fiber reinforced concrete at elevated temperatures. *Construction and Building Materials*, 313, 125437.

[15] Chen X, Li Q and Wang J (2020). *A Unified Sequence Labeling Model for Emotion Cause Pair Extraction*. In: Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020). Barcelona, Spain: International Committee on Computational Linguistics, pp. 208–218.

[16] Craswell N and Robertson S (2009). Average Precision at n. *Encyclopedia of Database Systems*, pp. 193–194.

[17] Demenech LM, Oliveira AT, Neiva-Silva L and Dumith SC (2021). Prevalence of anxiety, depression and suicidal behaviors among Brazilian undergraduate students: A systematic review and meta-analysis. *Journal of Affective Disorders*, 282, pp. 147–159.

[18] Devlin J, Chang M, Lee K and Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv Preprint*, arXiv:1810.04805.

[19] Duarte D, El-Hagrassy MM, Couto TCE, Gurgel W, Fregni F and Correa H (2020). Male and Female Physician Suicidality: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, 77(6), pp. 587–597.

[20] Fleiss JL (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), pp. 378–382.

[21] Fujiwara K, Huang Y, Hori K, Nishioji K, Kobayashi M, Kamaguchi M and Kano M (2020) Over- and Under-sampling Approach for Extremely Imbalanced and Small Minority Data Problem in Health Record Analysis. *Frontiers in Public Health*, 8, pp. 178.

[22] Gaur M, Alambo A, Sain JP, Kursuncu U, Thirunarayan K, Kavuluru R, Sheth A, Welton R and Pathak J (2019). *Knowledge-Aware Assessment of Severity of Suicide Risk for Early Intervention*. In: Proceedings of The World Wide Web Conference (WWW'19). New York, USA: Association for Computing Machinery, pp. 514–525.

[23] Gaur M, Aribandi V, Alambo A, Kursuncu U, Thirunarayan K, Beich J, Pathak J and Sheth A (2021). Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. *PLoS One*, 16(5), e0250448.

[24] Gee BL, Han J, Benassi H and Batterham PJ (2020). Suicidal thoughts, suicidal behaviours and self-harm in daily life: A systematic review of ecological momentary assessment studies. *Digital Health*, 6, 2055207620963958.

[25] Ghosh A and Lan A (2021). Do We Really Need Gold Samples for Sample Weighting Under Label Noise. *ArXiv Preprint*, arXiv:2104.09045.

[26] Gkotsis G, Velupillai S, Oellrich A, Dean H, Liakata M, Dutta R (2016). *Don't Let Notes Be Misunderstood: A Negation Detection Method for Assessing Risk of Suicide in Mental Health Records*. In: Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology. San Diego, USA: Association for Computational Linguistics, pp. 95–105.

[27] Grimmond J, Kornhaber R, Visentin D, Cleary M (2019). A qualitative systematic review of experiences and perceptions of youth suicide. *PLOS ONE*, 14(6), pp. 1–25.

[28] Hassana R (1998). One Hundred Years of Emile Durkheim's Suicide: A Study in Sociology. *Australian & New Zealand Journal of Psychiatry*, 32(2), pp. 168–171.

[29] Hernández-Bello L, Hueso-Montoro C, Gómez-Urquiza JL, Cogollo-Milanés Z (2020). Prevalence and asociated factor for ideation and suicide attempt in adolescents: a systematic review. *Revista Española de Salud Pública*, 94, e202009094.

[30] Hill NTM, Robinson J, Pirkis J, Andriessen K, Krysinska K, Payne A, Boland A, Clarke A, Milner A, Witt K, Krohn S and Lampit A (2020). Association of suicidal behavior with exposure to suicide and suicide attempt: A systematic review and multilevel meta-analysis. *PLoS Med*, 17(3), e1003074.

[31] Howarth EJ, O'Connor DB, Panagioti M, Hodkinson A, Wilding S and Johnson J (2020). Are stressful life events prospectively associated with increased suicidal ideation and behaviour? A systematic review and meta-analysis. *Journal of Affective Disorders*, 266, pp. 731–742.

[32] Jafari H, Heidari M, Heidari S and Sayfouri N (2020). Risk Factors for Suicidal Behaviours after Natural Disasters: A Systematic Review. *Malaysian Journal of Medical Sciences*, 27(3), pp. 20–33.

[33] Ji S, Yu CP, Fung S, Pan S, Long G and Cong G (2018). Supervised Learning for Suicidal Ideation Detection in Online User Content. *Complexity*, 6157249.

[34] 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), pp. 214–226.

[35] Kahil K, Cheaito MA, Hayek RE, Nofal M, Halabi SE, Kudva KG, Pereira-Sanchez V and Hayeka SE (2021). Suicide during COVID-19 and other major international respiratory outbreaks: A systematic review. *Asian Journal of Psychiatry*, 56, 102509.

[36] Kour H and Gupta MK (2022). *Depression and Suicide Prediction Using Natural Language Processing and Machine Learning*. In: Proceedings of the International Conference on Cyber Security, Privacy and Networking (ICSPN 2021). India: Springer Nature Singapore, pp. 117-128.

[37] Lau MA, Segal ZV and Williams JM (2004). Teasdale's differential activation hypothesis: implications for mechanisms of depressive relapse and suicidal behaviour. *Behaviour Research and Therapy*, 42(9), pp. 1001–1017.

[38] Li Q, Yan Z, Li J, Yang Z, Lin Z, Leong HV, Chen L and Yu NX (2021). *Event Cube for Suicidal Event Analysis: A Case Study*. In: Proceedings of Web Information Systems Engineering – WISE 2021. Melbourne, Australia: Springer, Cham, pp. 512–526.

[39] Li Q, Ma Y and Yang Z (2017). Event Cube – A Conceptual Framework for Event Modeling and Analysis. In: Proceedings of Web Information Systems Engineering – WISE 2017. Puschino, Russia: Springer, Cham, pp. 499–515.

[40] Loshchilov I and Hutter F (2019). Decoupled Weight Decay Regularization. *ArXiv Preprint*, arXiv:1711.05101.

[41] Méndez-Bustos P, Calati R, Rubio-Ramírez F, Olié E, Courtet P and Lopez-Castroman J (2019). Effectiveness of Psychotherapy on Suicidal Risk: A Systematic Review of Observational Studies. *Frontiers in Psychology*, 10, 277.

[42] Miranda-Mendizabal A, Castellví P, Parés-Badell O, Alayo I, Almenara J, Alonso I, Blasco MJ, Cebrià A, Gabilondo A, Gili M, Lagares C, Piqueras JA, Rodríguez-Jiménez T, Rodríguez-Marín J, Roca M, Soto-Sanz V, Vilagut G and Alonso J (2019). Gender differences in suicidal behavior in adolescents and young adults: systematic review and meta-analysis of longitudinal studies. *International Journal of Public Health*, 64(2), pp. 265–283.

[43] Norval C and Henderson T (2017). Contextual Consent: Ethical Mining of Social Media for Health Research. *ArXiv Preprint*, arXiv: 1701.07765.

[44] O'Connor RC and Kirtley OJ (2018). The integrated motivational-volitional model of suicidal behaviour. *Philosophical Transactions of the Royal Society B*, 373, 20170268.

[45] Ong MS, Lakoma M, Bhosrekar SG, Hickok J, McLean L, Murphy M, Poland RE, Purtell N and Ross-Degnan D (2021). Risk factors for suicide attempt in children, adolescents, and young adults hospitalized for mental health disorders. *Child and Adolescent Mental Health journal*, 26(2), pp. 134–142.

[46] Opitz J and Burst S (2021). Macro F1 and Macro F1. *ArXiv Preprint*, arXiv:1911.03347.

[47] Pengpid S and Peltzer K (2020). Suicide attempt and associated factors among in-school adolescents in Mozambique. *Journal of Psychology in Africa*, 30(2), pp. 130–134.

[48] Rabani ST, Khan QR and Ud Din Khanday AM (2020). Detection of suicidal ideation on Twitter using machine learning & ensemble approaches. *Baghdad Science Journal*, 17(4), pp. 1328-1339

[49] Richardson C, Robb KA and O'Connor RC (2021). A systematic review of suicidal behaviour in men: A narrative synthesis of risk factors. *Social Science & Medicine*, 276, 113831.

[50] Rudd MD, Berman AL, Joiner TE, Nock MK, Silverman MM, Mandrusiak M, Orden KV and Witte T (2006). Warning signs for suicide: theory, research, and clinical applications. *Suicide and Life-Threatening Behavior*, 36(3), pp. 255–262.

[51] Sawhney R, Manchanda P, Singh R and Aggarwal S (2018). *A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets*. In: Proceedings of ACL 2018, Student Research Workshop. Melbourne, Australia: Association for Computational Linguistics, pp. 91–98.

[52] 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. Brussels, Belgium: Association for Computational Linguistics, pp. 167–175.

[53] Sawhney R, Joshi H, Gandhi S and Shah RR (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). Online: Association for Computational Linguistics, pp. 7685–7697.

[54] Sawhney R, Joshi H, Flek L and Shah RR (2021). *PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media*. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, pp. 2415–2428.

[55] Shing H, Nair S, Zirikly A, Friedenberg M, Daumé H 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. New Orleans, LA: Association for Computational Linguistics, pp. 25–36.

[56] Shing H, Resnik P and Oard D (2020). *A Prioritization Model for Suicidality Risk Assessment*. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, pp. 8124–8137.

[57] Sinha PP, Mishra R, Sawhney R, Mahata D, Shah RR and Liu H (2019). *#suicidal - A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter*. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). New York, NY, USA: Association for Computing Machinery, pp. 941–950.

[58] Surace T, Fusar-Poli L, Vozza L, Cavone V, Arcidiacono C, Mammano R, Basile L, Rodolico A, Bisicchia P, Caponnetto P, Signorelli MS and Aguglia E (2021). Lifetime prevalence of suicidal ideation and suicidal behaviors in gender non-conforming youths: a meta-analysis. *European Child & Adolescent Psychiatry*, 30(8), pp. 1147–1161.

[59] Vioulès MJ, Moulahi B, Azé J and Bringay S (2018). Detection of suicide-related posts in Twitter data streams. *IBM Journal of Research and Development*, 62(1), p. 7:1-7:12.

[60] Wenzel A and Beck AT (2008). A cognitive model of suicidal behavior: Theory and treatment. *Applied and Preventive Psychology*, 12(4), pp. 189–201.

[61] Werbart Törnblom A, Sorjonen K, Runeson B and Rydelius P (2020). Who Is at Risk of Dying Young from Suicide and Sudden Violent Death? Common and Specific Risk Factors among Children, Adolescents, and Young Adults. *Suicide and Life-Threatening Behavior*, 50(4), pp. 757–777.

[62] Zhang Z and Sabuncu MR (2018). Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *ArXiv Preprint*, arXiv:1805.07836.

[63] Zirikly A, Resnik P, Uzuner Ö 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. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 24–33.