Social Science & Medicine 283 (2021) 114176

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Detecting suicide risk using knowledge-aware natural language processing and counseling service data

Zhongzhi Xu [a,b,1], Yucan Xu [b,1], Florence Cheung [b], Mabel Cheng [b], Daniel Lung [b], Yik Wa Law [b], Byron Chiang [b], Qingpeng Zhang [a,**], Paul S.F. Yip [b,*]

[a] School of Data Science, City University of Hong Kong, Hong Kong Special Administrative Region
[b] Hong Kong Jockey Club Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong Special Administrative Region

ARTICLE INFO

Keywords:
Online counseling services
Suicide prevention
Natural language processing
Knowledge graph
Artificial intelligence

ABSTRACT

Rationale: Detecting users at risk of suicide in text-based counseling services is essential to ensure that at-risk individuals are flagged and prioritized.
Objective: The objective of this study is to develop a domain knowledge-aware risk assessment (KARA) model to improve our ability of suicide detection in online counseling systems.
Methods: We obtained the largest known de-identified dataset from an emotional support system established in Hong Kong, comprising 5682 Cantonese conversations between help-seekers and counselors. Of those, 682 conversations disclosed crisis intentions of suicide. We constructed a suicide-knowledge graph, representing suicide-related domain knowledge as a computer-processible graph. Such knowledge graph was embedded into a deep learning model to improve its ability to identify help-seekers in crisis. As the baseline, a standard NLP model was applied to the same task. 80% of the study samples were randomly sampled to train model parameters. The remaining 20% were used for model validation. Evaluation metrics including precision, recall, and c-statistic were reported.
Results: Both KARA and the baseline achieved high precision (0.984 and 0.951, shown in Table 2) and high recall (0.942 and 0.947) towards non-crisis cases. For crisis cases, however, KARA model achieved a much higher recall than the baseline (0.870 vs 0.791). The c-statistics of KARA and the baseline were 0.815 and 0.760, respectively.
Conclusion: KARA significantly outperformed standard NLP models, demonstrating good translational value and clinical relevance.

1. Introduction

Suicide prevention has become an increasingly significant challenge worldwide (Fahey et al., 2020; Kirtley and O'Connor, 2020; Lutter et al., 2020; Naghavi, 2019). Approximately every 40 s globally, one person dies by suicide, which accounts for 1.5% of all deaths (World Health Organization, 2019). Most of the time, however, suicide is not inevitable if the risk can be identified in time. A person in crisis needs immediate support, in the form of medical emergency management, intervention or therapy (Kelly et al., 2008; McClatchey et al., 2019). Therefore, it is critical to identify at-risk persons in a timely fashion so that lives can be saved with appropriate support and intervention (Kelly et al., 2008; McClatchey et al., 2019). Identifying at-risk persons is especially important when counseling occurs online, where the remote nature of the engagement might impact counselors' ability to administer a sufficiently thorough assessment and timely triaging (Gilmore and Ward-Ciesielski, 2019).

Online counseling services have been rapidly growing since they commenced in the 1990s (Alao et al., 1999; Bantilan et al., 2021; Jashinsky et al., 2014; Kiesler, 1997). These services offer an increasingly-important way of detecting and preventing suicide in vulnerable people (Bantilan et al., 2021). Not only are online counseling services convenient, accessible, and timely for people in urgent need of help, but the anonymous nature of the Internet can increase

* Corresponding author. Hong Kong Jockey Club Centre for Suicide Research and Prevention, The University of Hong Kong, Hong Kong Special Administrative Region.
** Corresponding author. School of Data Science, City University of Hong Kong, Hong Kong Special Administrative Region.
E-mail addresses: qingpeng.zhang@cityu.edu.hk (Q. Zhang), sfpyip@hku.hk (P.S.F. Yip).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.socscimed.2021.114176
Received 29 June 2020; Received in revised form 4 June 2021; Accepted 21 June 2021
Available online 25 June 2021

help-seekers' comfort and autonomy by enhancing their sense of control (Chan, 2020). The role and importance of online emotional support services have become particularly important during the Covid-19 pandemic, which has largely disrupted traditional face-to-face services (Yip and Chau, 2020). Online engagement between counselors and help-seekers also provides an ideal source of psychotherapy data for natural language processing (NLP) models, as all interactions can be captured and processed automatically through the digital interface.

Data mining of online text-based counseling messages offers an important and promising way of calculating the likelihood of suicide. Although, there has been little research into this topic. Literature related to online text-based counseling services has typically focused on one or more of four aspects: counseling skills and counselor training (Kraus, 2004; Mallen et al., 2005); post-session referral and interventions (Krysinska and De Leo, 2007; Mallen et al., 2005); legal and ethical issues (Kraus, 2004; Mallen et al., 2005; McVeigh and Heward-Belle, 2020); and assessment of therapeutic outcomes (Chan, 2020; Hanley and Reynolds, 2009). To our knowledge, there is only one recent study that has investigated the automatic detection of suicide risk in online counseling services using machine learning algorithms. The authors shared our concern that suicide detection systems for telemedicine psychotherapy settings are important but as yet under-researched (Bantilan et al., 2021).

This knowledge gap may reflect one or more of the following reasons. Firstly, recruiting and training qualified counselors and social workers to annotate suicide risk level in counseling text dataset is costly and time-consuming. Secondly, data on high-risk cases are rare, thus true positive data are insufficient for deep learning models to learn suicidal language patterns. This low-base-rate problem is a common challenge in suicide and self-harm studies (Bantilan et al., 2021). Thirdly, suicide detection and sentiment analysis are different. Relatively speaking, suicide detection is highly domain specific. Consequently, standard NLP models designed to understand general linguistic patterns (Cheng et al., 2015; Gkotsis et al., 2016; Pavalanathan and De Choudhury, 2015) may not be the most appropriate to the specific needs of identifying suicide risk from social media posts.

We attempted to tackle these issues by making use of the recently-developed knowledge-aware models in the research field of computer science (Xu et al, 2020, 2021). More concretely, we proposed and tested a novel suicide knowledge graph-powered NLP approach which incorporates a knowledge graph into a standard deep learning model. The suicide knowledge graph was constructed by licensed psychologists and social workers. In this way, expert domain knowledge was translated into structured data which is processable by computers. To preface our results, we found that the proposed model incorporating expert domain knowledge to compensate for the relatively rare suicide linguistic patterns significantly outperformed baselines (c-statistic, 0.82 versus 0.76), providing a potential solution over the challenge of finding a needle in a stack of hay.

2. Methods

2.1. Dataset and case identification

This study analyzed secondary anonymous data obtained from online conversations recorded on OpenUp, which is a 24-h online text-based platform designed specifically to cater to the needs of youngsters experiencing emotional distress from a range of concerns, such as family issues, interpersonal relationships, and academic stress (Yip et al., 2020). Youngsters can use this platform anonymously to express their distress, whilst feeling a sense of companionship and self-determination. Up to now, 224 counselors and trained volunteers so far have been engaged in this service. Because of the level of demand, counselors may have to chat with more than one youngster simultaneously. This sort of one-to-many arrangement is also very different from the traditional counseling session between counselor and the user on a one-to-one basis. The good practice model of this one-to-many model has yet to be developed.

Fig. 1 shows a fictitious conversation between a help-seeker and a counselor in Cantonese (with English translation). The full dataset contains over 22,000 conversations collated in the year 2019. The suicide risk level of the help-seeker was annotated in each conversation by experienced social workers and counselors. They were given clear instructions on what kind of content is associated with particular crisis levels, via comprehensive materials including typical examples, common mistakes, and helpful tips. As a result, each conversation was classified into one of the two ordinal categories: crisis and non-crisis (see the definitions in Table S1). For each conversation, three social workers and/or counselors independently performed the annotation task, producing a high inter-annotator agreement rate (Krippendorff's α = 95.6%; Hayes and Krippendorff, 2007). Conversations with ambiguous risk levels were not included in the analysis of this study.

For this study, we extracted a subset of the full dataset, identifying 5682 conversations that had at least ten messages exchanged between the help-seeker and the counselor (ping-pong count $\geq$ 10). Of those, 682 conversations were designated crisis level, whilst the remainder were with non-crisis suicide risk. We formulated the suicide risk detection task as a binary classification problem (crisis level vs non-crisis level).

2.2. Patient and public involvement

This study was secondary anonymous data analysis. It was not appropriate or possible to involve patients or the public in the design, or conduct, or reporting, or dissemination of our research.

2.3. The risk assessment model

We developed a novel NLP model, namely the *knowledge-aware risk assessment (KARA)* model, to detect the risk of suicide through mining conversations between help-seekers and counselors. We begin with a brief description of the overall picture, then introduce its technical details.

Fig. 2 presents the overall architecture of the model. In general, KARA is constituted by two components: the *knowledge encoder* component and the *conversation encoder* component. We relied on the knowledge encoder component to learn representations of suicide-related domain knowledge. We then fed the encoded domain knowledge into the conversation encoder component which mined conversations, with the encoded domain knowledge, to assess the risk of suicide.

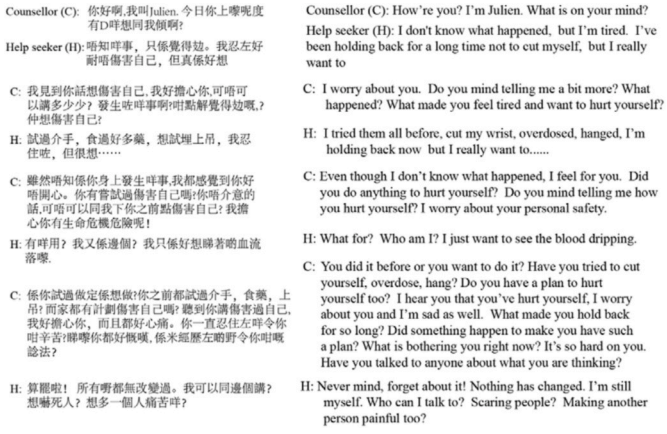

Fig. 1. A fictitious example of the conversation between a help-seeker and a counselor (with English translation).

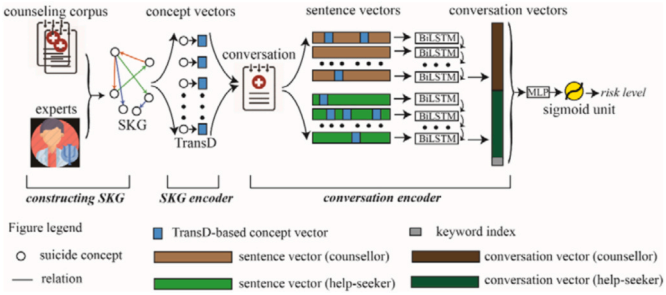

Fig. 2. The Architecture of KARA. KARA consists of two components: a knowledge graph encoder and a conversation encoder. The knowledge graph encoder takes the knowledge graph as input, and outputs TransD-based concept vectors (blue boxes). The conversation encoder takes a conversation as input, and outputs a conversation vector. Note that counselor's message and the help-seeker's message in one conversation are embedded separately because the two roles' linguistic patterns are usually different. Brown sentence vectors and the dark brown conversation vector were derived from counselor's message, green sentence vectors and the dark green conversation vector were derived from help-seeker's message. In the sense of incorporating TansD-based concept vectors into sentence vectors, KARA is knowledge-aware. The conversation vector is then fed into a Multi-Layer Perceptron (MLP) module and a sigmoid unit to output the probability score of a help-seeker being in crisis. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

2.4. Suicide knowledge graph

We formulated domain knowledge as a *knowledge graph* such that the data could be processed by computers. A knowledge graph was constructed by structured data in the form of knowledge triples (*head concept, relation, tail concept*), where *relation* represents the relationship between the head and tail concepts. In the context of suicide, a group of experienced counselors and social workers were recruited to create a suicide knowledge graph (SKG). In the SKG, a node represents a suicide-related key concept. Examples of concepts (i.e., nodes) include sleeping pills, knife, overdosage, hanging, self-harm, cutting, bleeding, rooftop, parents, lose one's job, and so forth. Edges in SKG represent the semantic relationship between the two corresponding concepts. For example, "*break up* (分手)" could be the cause of "*unhappy* (唔開心)", thus there is a labeled edge ("*because of*") pointing at "*unhappy*" from "*break up*". As a result, SKG consists of a set of *concept-relation-concept* triples, covering the 148 most prevalent and important concepts in counseling conversations. Refer to SI-*II-B-Concept selection rule* for the inclusion method of concepts.

We considered six types of semantic relations, namely "*may be because of*", "*similar to*", "*may lead to*", "*same as*", "*may relate to*" and "*opposite to*" in SKG. As a result, we obtained 921 knowledge triples. We explained why we focused on these relations in SI-*II-C-Relation selection rule*. Table S2 shows the distribution of the number of each relation. Fig. 3 demonstrates the SKG.

2.5. Encoding the suicide knowledge graph

The core idea of the knowledge graph embedding model is that every semantic relationship is regarded as a translation of the embedding space. For a knowledge triple (h, r,t), the embedding of concept $\boldsymbol{h}$ is close to the embedding of concept $\boldsymbol{t}$ by adding the embedding of semantic relationship $\boldsymbol{r}$, that is $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$ (Fig. 4). In the KARA model, in particular, a popular graph embedding method called TransD was adopted to derive *SKG-based concept vectors*. Concepts not included in SKG were embedded by word2vec, a popular language pre-training model, to derive the *word2vec-based concept vectors*. (See SI-*II-D* for more details).

To demonstrate the efficacy of word2vec, we calculated the semantic similarity between "jumping" and other words in the corpus and listed the most similar words as an example. Following the same procedure, words similar to "cutting" were also listed as another example. To demonstrate the efficacy of TransD, we presented eight randomly sampled SKG-based concept vectors in three-dimensional space.

2.6. Encoding the conversation

The conversation encoder component encoded a conversation into a vector (called a *conversation vector*). Encoding a conversation requires two steps. The first was concatenating concept vectors in a sentence, obtaining the sentence vector. Note that SKG-based concept vectors (if existed) in lieu of word2vec-based concept vectors were adopted. In this sense, the proposed model is knowledge aware. The second was feeding sentence vectors of the counselor (colored in brown in Fig. 2) and the help-seeker (colored in green in Fig. 2) into two separate Bidirectional Long Short-Term Memory (BiLSTM) neural network modules to generate the counselor-message vector and help-seeker-message vector (See SI-*II-E* for details about BiLSTM). Such two vectors were then concatenated into one vector, namely a *conversation vector*. We also included an indicator (0 or 1) at the end of the conversation vector representing the appearance of concepts related to the means of suicide (such as "rope" and "knife") if these appeared in the help-seeker's message (grey box in Fig. 2). Lastly, feeding the conversation vector into an MLP neural network module and a sigmoid unit, the KARA model output the probability score of a user at the crisis level of suicide.

2.7. Model evaluation

We compared KARA with a baseline BiLSTM deep learning model, which is essentially the KARA model without SKG (that is, without domain knowledge). Such a baseline is the standard approach in text analysis studies (Aziz Sharfuddin et al., 2018). We randomly selected 80% of the study samples to train KARA and the baseline. The remaining 20% were used for model validation. To conduct a comprehensive evaluation of model performance, it was necessary to consider multiple metrics from different perspectives. In this study, we reported the precision, recall (also known as sensitivity), and c-statistic (also known as the area under the receiver operating characteristic curve (ROC-AUC)). Negative cases were conversations with non-crisis risk level. Positive cases were conversations with crisis risk level. A trade-off threshold of 0.5 between precision and recall was used in all statistical analyses. We also tested the KARA model in a real-time scenario by continuously monitoring a fictitious conversation. To check the robustness of the KARA model, we performed several sensitivity analyses (reported in SI-*III-A*).

3. Results

3.1. Concept vectors and the conversation vector

3.1.1. Word2vec-based concept vectors

Given word2vec-based concept vectors, Table 1 lists concepts similar to "jumping" and "cutting" based on cosine similarity. Concepts in the first column correspond to the top-10 words similar to "jumping". Values for ranking in the second column refer to the corresponding semantic similarities. Concepts similar to "cutting" and corresponding similarities were also shown in Table 1 (third and fourth columns). It can be observed that word2vec-based concept vectors can capture the semantic relations between concepts.

3.1.2. SKG-based concept vectors

Fig. 5 demonstrates SKG-based concept vectors of eight knowledge triples in three-dimensional space, containing thirteen concepts and two types of relations. It can be observed that semantic relations of concepts were captured by SKG-based concept vectors in this three-dimensional space: Semantic relations of same relation type (distinguished by line

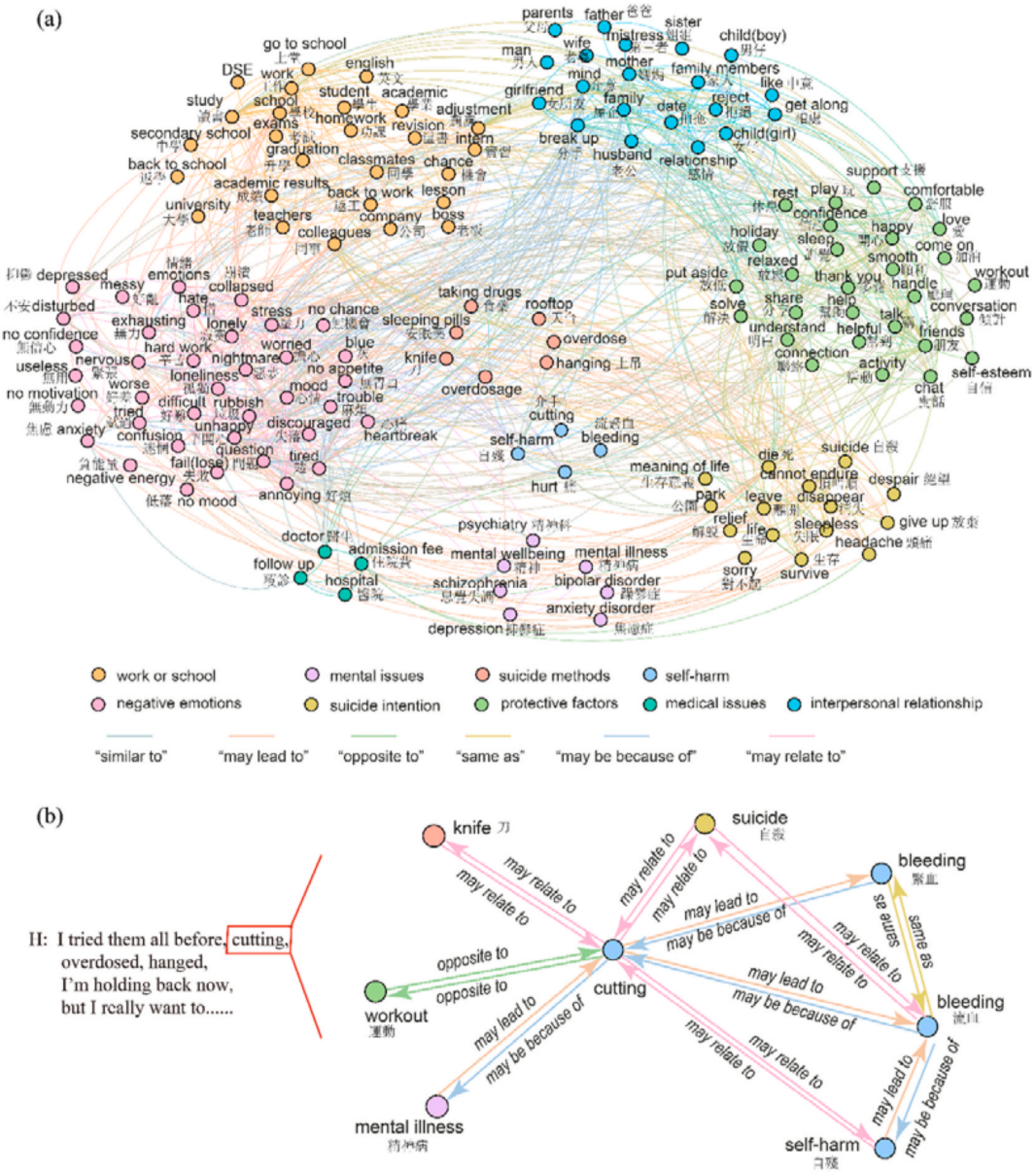

Fig. 3. (a) The suicide knowledge graph in English. The color of node represents different categories of concepts. One concept may belong to two or more categories. In such case, one category is randomly selected for clearer visualization. The color of edge represents different types of semantic relations. English words such as "overdose", "overdosage" and "DSE" (Diploma of Secondary Education) are usually used directly in Cantonese conversations. Such words do not have corresponding Cantonese. (b) **An example demonstrating concepts adjacent to "cutting".** Note that, as shown in the fictitious example, concepts semantically related to "cutting" may not occur explicitly in the conversation. We hypothesized that incorporating implicit semantic contexts could enhance the model's accuracy in evaluating the risk of suicide. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

color) were relatively parallel. We use SKG-based concept vectors with eight dimensions in model training to include more information. We were not able to visualize such a dimension.

3.1.3. The conversation vector

We presented three-dimensional conversation vectors of 200 randomly sampled positive cases (100) and counterparts (100) in Fig. 6, where red circles represent crisis conversations and blue circles represent non-crisis conversations. The observation is that three-dimensional conversation vectors roughly distinguished the two groups of conversations. In model training, the dimension of the vector is much larger (128 dimensions) to better distinguish the two groups. Again, presenting such a dimension in a plain is not feasible.

3.2. Model performance

Both KARA and the baseline model achieved high precision (0.984 and 0.951, shown in Table 2) towards the negative cases ($\frac{TN}{TN+FN}$) and high recall (0.942 and 0.947, shown in Table 2) towards the negative cases ($\frac{TN}{TN+FP}$). This outcome was expected because of the imbalanced distribution of study samples (skewed to the negative).

For positive cases, which were clinically important as they referred to users at risk, the precision ($\frac{TP}{TP+FP}$) of the KARA model was higher than that of the baseline (0.649 vs 0.633, shown in Table 2). More importantly, the KARA model achieved a much higher recall for positive cases ($\frac{TP}{TP+FN}$) than the baseline (0.870 vs 0.791, shown in Table 2). This encouraging result indicated that KARA reflected 87% of the high-risk cases, 10% more than the baseline BiLSTM model.

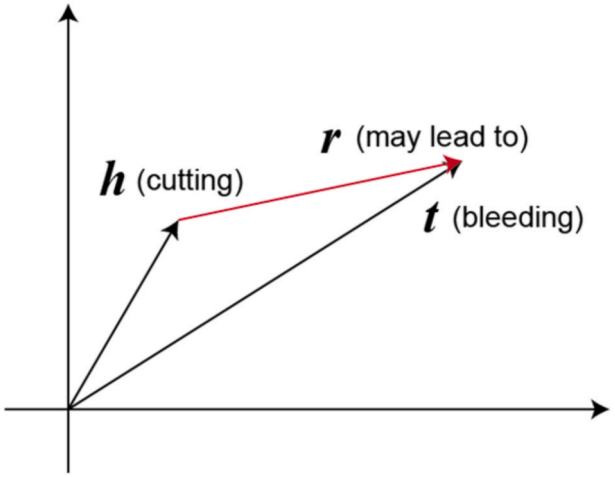

Fig. 4. An Example Illustrating the Basic Idea of Knowledge Graph Embedding. Note. For example, "cutting", "may lead to", "bleeding" are a knowledge triple (h, r, t). The aim is to find appropriate vectors $\boldsymbol{h}$, $\boldsymbol{r}$ and $\boldsymbol{t}$ to represent h, r and t in a knowledge triple (h, r, t) so that $\boldsymbol{h} + \boldsymbol{r} \approx \boldsymbol{t}$.

Table 1
Two examples of the word2vec-based concept vector similarity.

concepts affinitive to "jumping" ("跳樓")	cosine similarity	concepts affinitive to "cutting" ("介手")	cosine similarity
jumping into the ocean (跳海)	0.798	self-harm (自殘)	0.731
jumping (跳落去)	0.758	cutting (割)	0.635
die (死)	0.736	bleeding (流血)	0.610
hanging (吊頸)	0.735	after cutting (介完)	0.601
charcoal burning (燒炭)	0.730	cutting (界手)	0.587
suicide (自殺)	0.710	run away from home (離家出走)	0.565
rushing out (衝出)	0.696	flowing (流緊)	0.560
wrist cutting (割脈)	0.694	jumping (跳樓)	0.553
self-harm (自殘)	0.671	bleeding (緊血)	0.540
suicide (輕生)	0.665	wrist (手腕)	0.536

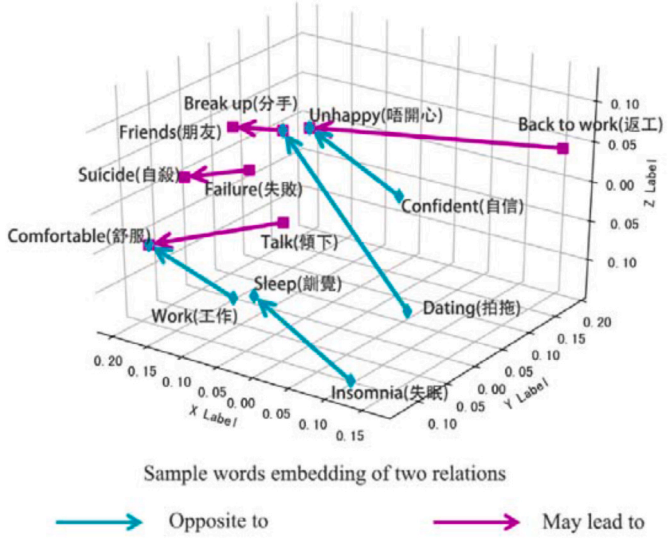

Fig. 5. An illustration of embeddings of eight knowledge triples in three-dimensional space.

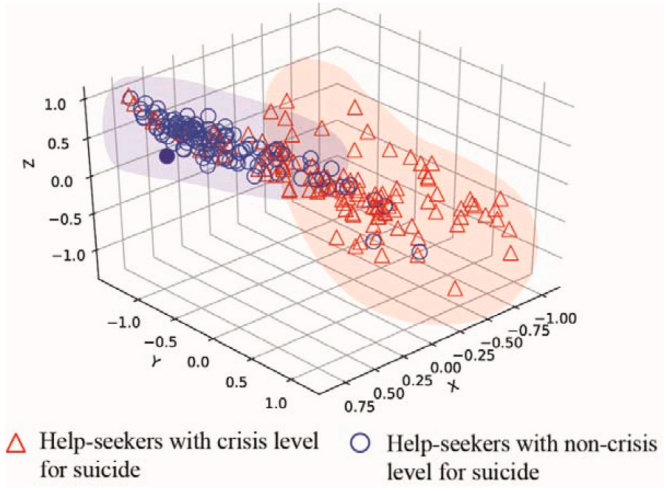

Fig. 6. An illustration of 200 conversation vectors in three-dimensional space.

Table 2
The performance of the baseline BiLSTM model and the KARA model.

baseline BiLSTM	**precision (*SD*)**	**recall (*SD*)**	**c-statistic (*SD*)**
non-crisis conversations	0.951 (0.006)	0.947 (0.006)	0.760 (0.007)
crisis conversations	0.633 (0.025)	0.791 (0.010)	
KARA	**precision (*SD*)**	**recall (*SD*)**	**c-statistic (*SD*)**
non-crisis conversations	0.984 (0.004)	0.942 (0.006)	0.815 (0.006)
crisis conversations	0.649 (0.028)	0.870 (0.017)	

The c-statistics of KARA and the baseline were 0.815 and 0.760, respectively, further demonstrating the advantage of the proposed model. Sensitivity analysis demonstrated the robustness of this advantage (Fig. S1 and Table S3).

Precision-Recall curves summarized the trade-off between the true positive rate and the positive predictive value for a classification model using different probability thresholds. Fig. 7A and B use precision-recall curves to illustrate the variance of recall and precision.

3.3. Real-Time Evaluation of Suicide Risk

We also tested the effectiveness of the proposed KARA model in a real-time scenario. As Fig. 8 shows, the model provided real-time feedback (normalized risk level of suicide) to counselors. In the meantime, a counselor was also engaged in evaluating how the risk levels varied with the progress of the conversation. We observed that the trained model performed well for real-time conversations as well.

4. Discussion

To our knowledge, KARA was the first comprehensive deep learning model that incorporated the linguistic patterns from both counseling texts and external domain knowledge. As it incorporated two-side information, it should enable higher-resolution modeling for detecting the risk of suicide. Using large-scale counseling data with high-quality annotations, we conducted experiments to validate such assumptions. Extensive experiments showed that KARA significantly outperformed the baseline model that did not take into consideration domain knowledge. The 10% gap in covering cases in crisis provided strong evidence that incorporating the domain knowledge can significantly improve the capability of deep learning models in identifying those in crisis. The present study, along with a very recent study (Bantilan et al., 2021), were the pioneer studies in this line of research. The present study further confirmed that it is both possible and helpful to deploy an

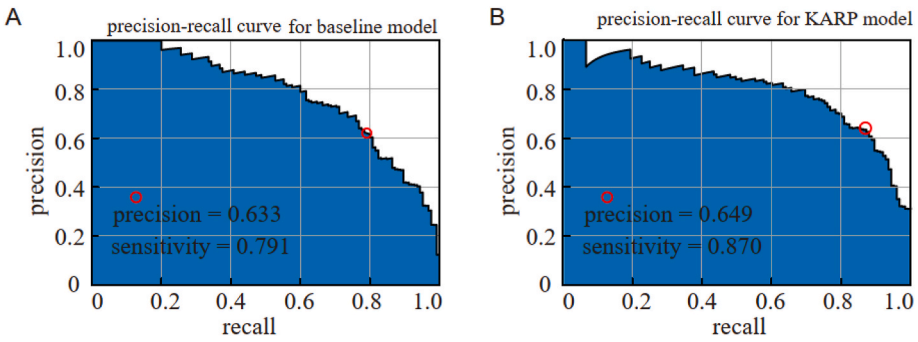

Fig. 7. The Precision-Recall Curves of (a) the Baseline Model and (b) KARA. Red circles demonstrate where a threshold of 0.5 located. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

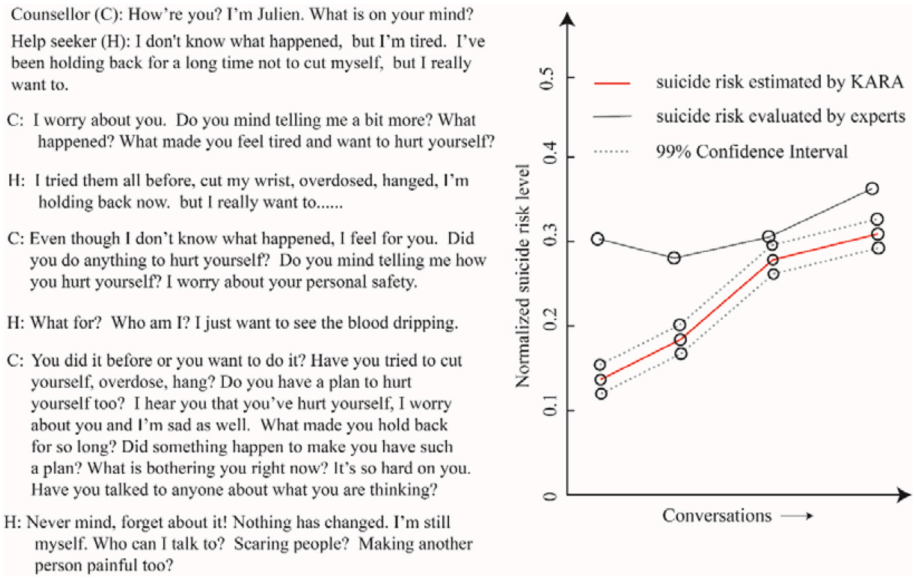

Fig. 8. An Example of the Real-Time Evaluation of Suicide Risk. The experiment was conducted 50 times. Black and red solid lines represent the variance of risk with time, evaluated by human and KARA, respectively. Dashed lines indicate the 99% CI. The conversation is fictive. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

accurate, passive, and automatic suicide risk detection model for alerting counselors to the presence of potential risk in a user's content during the engagement process. The proposed model did not mean to dictate the judgment of the counselor but instead provide alert and decision support. Another contribution compared to the status quo (Bantilan et al., 2021) arose from the proposed novel KARA model. It incorporated knowledge from experts which is in the form of a knowledge graph that depicts the semantic relationships between key concepts to complement suicide linguistic patterns such that the model performances were improved.

The idea of using domain knowledge to complement model training is not uncommon, especially in computer science (Wang et al., 2018; Wang et al., 2019a, 2019b; Xu et al., 2019a). However, this framework has been surprisingly underutilized by healthcare researchers. A possible explanation was the lack of available knowledge graph. In this study, the expert-generated SKG provided generic knowledge for suicide-related text analysis tasks. We have opened both the English and Cantonese versions of the SKG to the public on the GitHub repository.

The precision-recall curves illustrated the variance of precision and recall. We found that if we sacrificed precision and reduce it to 0.5 (meaning 50% false alarms), the recall increased to 0.9. In the context of suicide detection, we suggested that a high recall should be prioritized as long as the resources are sufficient. In this study, a trade-off threshold of 0.5 between precision and recall was used consistently in all statistical analyses to ensure fair comparisons between different experiments. In practice, health care professionals can adjust the threshold to achieve higher recall (or precision) according to contexts and circumstances (Xu et al, 2019b, 2020, 2019b).

The proposed methodology was also applicable to Online health communities (OHCs). OHCs are becoming popular for people to seek and offer emotional and psychological help. OHCs differed from counseling systems in two aspects: (a) OHCs were open-to-all, and (b) supporters in OHCs were usually not well-trained social workers or counselors. Despite the disparity, the proposed model can be extended to detect suicide risk in OHCs after recalibration over OHCs data under the premise that the data has been annotated by licensed social workers or counselors.

There are clear practical implications of this study. First, in real-world practice, counselor resources were usually insufficient to accommodate the huge number of help-seekers. The proposed approach can assist counselors in making an efficient, accurate, and instantaneous risk assessment, especially useful when they are engaged in chatting with multiple users. Secondly, the KARA model can assist counselors to prioritize those in crisis. This triage was essential because help-seekers who are at greatest risk should be prioritized in the counseling queue to be provided with sufficient attention. Shorter waiting time for people in need could prevent them from leaving the service prematurely; thus, it could potentially save lives. Finally, since the KARA model was trained

on thousands of conversations that were annotated by licensed experts, it was less subjective to individual bias. In this perspective, KARA was helpful to improve the quality of real-time risk assessment and post-session triage especially when volunteers and/or new recruits were involved.

Future research could consider the following promising directions. (a) Current literature has shown that the relationship between mental health status and media use might be conceptualized as a non-linear polynomial function (Scherr, 2018). Comparing the OpenUp users' profiles to the general population could add to a better understanding of mental health and social media use. (b) In recent years, as visually-driven social media platforms such as Instagram, Snapchat, or TikTok gained popularity, a challenge has arisen for frameworks that were developed for text-only social media platforms, including the one proposed in this study. To fill this research gap, in a recent study, Scherr et al. addressed the automatic detection of self-harm attempters through an analysis of images posted on Instagram (Scherr et al., 2020). At present, the OpenUp system only accepts text messages. We suggest that more studies are needed to explore the value of including videos and images in suicide detection. (c) The OpenUp platform explored in this study was primarily designed for young adults aged 11–35 years. Whether or not the proposed model works for other age groups has yet to be explored. This research direction is important because there is some evidence that suicidality increases use of social media platforms, especially among older users (Scherr and Reinemann, 2016). In addition, people of different age groups may have different linguistic patterns, thus the usefulness of the proposed model for different age groups and different linguistic patterns requires testing. Taken together, the platform can be further modified and enhanced to meet other age groups. The use of online emotional counseling has been gaining popularity and its service quality have yet to be established. As the mode of operation is very different from the traditional one-to-one counseling service, the good practice model has yet to be established. The appropriate use of artificial intelligence and big data analytics have yet to be fully explored and incorporated into enhancing the service. Certainly, the use of artificial intelligence and big data analytic is something future research should invest in to meet the changing need of our young people based on online emotional support.

4.1. Limitations

This study has two main limitations. First, the model was developed and trained to understand Cantonese linguistic patterns. Whether the model is similarly effective in other languages and cultural settings require further testing. Second, the proposed model is a black box because the inner decision process of neural networks is too complex to be translated into transparent rules. There will always be a tradeoff between model accuracy and model interpretability. As model accuracy is more important in the context of this study, we sacrificed interpretability to gain better performance.

5. Conclusion

We developed a novel model for detecting people at risk of suicide among online counseling service users aged 11–35. The proposed model incorporated domain knowledge into standard NLP models. The proposed model outperformed standard NLP models in various experiments, demonstrating good translational value and clinical relevance.

Declaration of competing interest

All authors declare no support from any organizations for the submitted work; no financial relationships with any organizations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.socscimed.2021.114176.

Funding source

This study is funded in part by the National Natural Science Foundation of China (NSFC) Grant Nos. 71972164 and 71672163, in part by the Health and Medical Research Fund Grant (HMRF) No. 16171991, in part by the Hong Kong Jockey Club Charity Trusts and in part by The Theme-Based Research Scheme of the Research Grants Council of Hong Kong Grant No. T32-102/14N.

Contributors

QZ and SFPY formulated the idea. ZX, YX, QZ and SFPY performed the literature review. ZX, YX, CB and LD developed the model and conducted the experiments. ZX, YX, CF, CM, QZ and SFPY analyzed and interpreted the results. QZ, ZX, YX, CF and SFPY wrote the article. All authors had full access to all data (including statistical reports and tables) in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

References

Alao, A.O., Yolles, J.C., Armenta, W., 1999. Cybersuicide: the Internet and suicide. Am. J. Psychiatr. 156, 1836–1837.
Aziz Sharfuddin, A., Nafis Tihami, M., Saiful Islam, M., 2018. A deep recurrent neural network with BiLSTM model for sentiment classification. In: 2018 Int. Conf. Bangla Speech Lang. Process. ICBSLP 2018.
Bantilan, N., Malgaroli, M., Ray, B., Hull, T.D., 2021. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. Psychother. Res. 31, 302–312.
Chan, G.H., 2020. A comparative analysis of online, offline, and integrated counseling among hidden youth in Hong Kong. Child. Youth Serv. Rev. 114, 105042.
Cheng, Q., Kwok, C.L., Zhu, T., Guan, L., Yip, P.S.F., 2015. Suicide communication on social media and its psychological mechanisms: an examination of Chinese microblog users. Int. J. Environ. Res. Publ. Health 12, 11506–11527.
Fahey, R.A., Boo, J., Ueda, M., 2020. Covariance in diurnal patterns of suicide-related expressions on Twitter and recorded suicide deaths. Soc. Sci. Med. 253, 112960.
Gilmore, A.K., Ward-Ciesielski, E.F., 2019. Perceived risks and use of psychotherapy via telemedicine for patients at risk for suicide. J. Telemed. Telecare 25, 59–63.
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, pp. 95–105.
Hanley, T., Reynolds, D., 2009. A review of the quantitative research into text-based therapy. Counsell. Psychol. Rev. 24, 4–13.
Hayes, A.F., Krippendorff, K., 2007. Answering the call for a standard reliability measure for coding data. Commun. Methods Meas. 1, 77–89.
Jashinsky, J., Burton, S.H., Hanson, C.L., West, J., Giraud-Carrier, C., Barnes, M.D., Argyle, T., 2014. Tracking suicide risk factors through Twitter in the US. Crisis 35, 51–59.
Kelly, C.M., Jorm, A.F., Kitchener, B.A., Langlands, R.L., 2008. Development of mental health first aid guidelines for suicidal ideation and behaviour: a Delphi study. BMC Psychiatr. 8, 1–10.
Kiesler, S., 1997. From the Couch to the Keyboard: Psychotherapy in Cyberspace. Cult. Internet 87–116.
Kirtley, O.J., O'Connor, R.C., 2020. Suicide prevention is everyone's business: challenges and opportunities for Google. Soc. Sci. Med. 262, 112691.
Kraus, R., 2004. Ethical and legal considerations for providers of mental health services online. In: Online Counseling: A Handbook for Mental Health Professionals, pp. 123–144.
Krysinska, K.E., De Leo, D., 2007. Telecommunications and suicide prevention: hopes and challenges for the new century. Omega J. Death Dying 55, 237–253.
Lutter, M., Roex, K.L.A., Tisch, D., 2020. Anomie or imitation? The Werther effect of celebrity suicides on suicide rates in 34 OECD countries, 1960–2014. Soc. Sci. Med. 246, 112755.
Mallen, M.J., Vogel, D.L., Rochlen, A.B., 2005. The practical aspects of online counseling: ethics, training, technology, and competency. Counsel. Psychol.
McClatchey, K., Murray, J., Chouliara, Z., Rowat, A., 2019. Protective factors of suicide and suicidal behavior relevant to emergency healthcare settings: a systematic review and narrative synthesis of post-2007 reviews. Arch. Suicide Res. 23, 411–427.
McVeigh, M.J., Heward-Belle, S., 2020. Necessary and good: a literature review exploring ethical issues for online counselling with children and young people who have experienced maltreatment. Child Aust. 45, 266–278.

Naghavi, M., 2019. Global, regional, and national burden of suicide mortality 1990 to 2016: systematic analysis for the Global Burden of Disease Study 2016. BMJ 364, l94.
Pavalanathan, U., De Choudhury, M., 2015. Identity management and mental health discourse in social media. In: WWW 2015 Companion - Proceedings of the 24th International Conference on World Wide Web, pp. 315–321.
Scherr, S., 2018. Traditional media use and depression in the general population: evidence for a non-linear relationship. Curr. Psychol.
Scherr, S., Arendt, F., Frissen, T., Oramas, M.J., 2020. Detecting intentional self-harm on Instagram: development, testing, and validation of an automatic image-recognition algorithm to discover cutting-related posts. Soc. Sci. Comput. Rev. 38, 673–685.
Scherr, S., Reinemann, C., 2016. First do no harm: cross-sectional and longitudinal evidence for the impact of individual suicidality on the use of online health forums and support groups. Comput. Hum. Behav. 61, 80–88.
Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M., 2019a. Exploring high-order user preference on the knowledge graph for recommender systems. In: ACM Transactions on Information Systems.
Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M., 2018. RippleNet: propagating user preferences on the knowledge graph for recommender systems. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 417–426.
Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S., 2019b. KGAT: knowledge graph attention network for recommendation. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 950–958.
World Health Organization, 2019. Global health estimates 2016: deaths by cause, age, sex, by country and by region, 2000–2016 [WWW Document]. URL (accessed 3.10.20). https://www.who.int/news-room/detail/09-09-2019-suicide-one-person-dies-every-40-seconds.
Xu, Z., Zhang, J., Zhang, Q., Xuan, Q., Yip, P.S.F., 2021. A comorbidity knowledge-aware model for disease prognostic prediction. IEEE Trans. Cybern. 1–11.
Xu, Z., Zhang, J., Zhang, Q., Yip, P.S.F., 2019a. Explainable learning for disease risk prediction based on comorbidity networks. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics. IEEE, pp. 814–818.
Xu, Z., Zhang, Q., Li, W., Li, M., Yip, P.S.F., 2019b. Individualized prediction of depressive disorder in the elderly: a multitask deep learning approach. Int. J. Med. Inf. 132, 103973.
Xu, Z., Zhang, Q., Yip, P.S.F., 2020. Predicting post-discharge self-harm incidents using disease comorbidity networks: a retrospective machine learning study. J. Affect. Disord. 277, 402–409.
Yip, P., Chan, W.L., Cheng, Q., Chow, S., Hsu, S.M., Law, Y.W., Lo, B., Ngai, K., Wong, K. Y., Xiong, C., Yeung, T.K., 2020. A 24-hour online youth emotional support: opportunities and challenges. Lancet Reg. Heal. - West. Pacific 4, 0–2.
Yip, P.S.F., Chau, P.H., 2020. Physical distancing and emotional closeness amidst COVID-19. Crisis 41, 153–155.