A Two-Stage Classification Chatbot for Suicidal Ideation Detection

Jin Xuan Chan, Sook-Ling Chua(✉), and Lee Kien Foo

Faculty of Computing and Informatics, Multimedia University, Cyberjaya, Selangor, Malaysia
slchua@mmu.edu.my

Abstract. Suicide remains one of the leading causes of death globally and is a serious public health problem. Compounded by the lack of mental health professionals and lack of access to mental health services, it is difficult for people with mental health issues to seek treatment. The advancements in artificial intelligence have led to the development of mental health digital solution, such as chatbots. A chatbot is a software application that simulates human conversations with users through text or voice interactions. Chatbots have been receiving increasing attention lately for its roles in providing alternative support and helping in filling the gaps in mental health care. Although there are many chatbots that are widely used for mental health, they are not designed to detect suicide risk. In this paper, a two-stage classification chatbot is proposed for suicidal ideation detection.

Keywords: Chatbot · Suicidal ideation detection · Deep learning · Mental health

1 Introduction

According to the World Health Organisation [1], mental health is defined as "a state of well-being in which an individual realises his or her own abilities, can cope with the normal stresses of life, can work productively and is able to make a contribution to his or her community." Poor mental health could lead to mental illnesses such as anxiety disorder and depression, which are serious conditions that may significantly affect the functioning of daily life.

Suicide is a common impact or risk for individuals with mental illness. About 83% of the suicide attempters had at least one mental illness [2]. Suicide is one of the leading causes of death worldwide. In 2019, Malaysia recorded a rate of 5.8 suicides per 100,000 population, which is approximately 5 deaths per day [3].

People with suicidal ideation and those that have taken their own life often have communicated their suicidal tendencies to mental health professionals [4]. However, due to the lack of mental health professionals and compounded by the lack of access to mental health services, it is indeed challenging for people seeking mental health treatment. Such barriers may lead to a higher incidence of mental health issues and increase in suicide.

Current intervention requires mental health professionals in screening individuals for suicide risk. However, it can be difficult as some individuals may attempt to conceal

S.-C. Haw and K. Sonai Muthu (Eds.): CITIC 2022, 10, pp. 405–412, 2022.
https://doi.org/10.2991/978-94-6463-094-7_31

suicidal ideation during clinical setting. This greatly impact on missed screening and prevent early intervention.

The advancements in digital technology and artificial intelligence (AI) have led to the development of chatbots, which seen as a viable method to supplement the traditional mental health professionals. Chatbot is an AI software that can mimic human-like behaviour to interact and conduct conversations with human users either through voice or text communication [5]. Studies have shown that chatbots can be an alternative and additional mental health support particularly in situation where people are feeling embarrass or discomfort in disclosing their mental health problems to a therapist [6]. Although there are a number of widely deployed chatbots for use in mental health care provision (e.g., Woebot, Wysa, etc.), these chatbots are mainly to support those with anxiety or depression and help users to improve their mental health. These chatbots are not designed to identify individuals who may be at risk for suicide.

The main research question that this paper aims to address is "how can a chatbot be designed to detect suicide risk from text conversation". The challenge in creating a chatbot lies in the ability of the chatbot to initiate conversation, provide appropriate response based on user's input and direct them to the mental health professionals, when necessary. In this paper, we propose a two-stage classification chatbot for suicidal ideation detection.

2 Related Work

This section reviews the related work in suicidal ideation detection. Earlier studies in suicidal ideation detection applied sentiment analysis to determine user's emotion on social media. In the work of [7], they built a suicide dictionary and calculate the semantic similarity for sentiment classification. In [8], they applied sentiment analysis on Tweet-based features obtained from user's profile and followees' information. However, only considering the sentiment of a post may not be effective to detect suicidal ideation.

Other studies attempt to extract features related to suicide by generating an n-gram language structure, which is then used to train supervised learning models for suicide ideation detection [9, 10]. Although n-gram has shown a number of success in text classification, it does not capture the long range dependencies among the words in text.

There are studies that attempt to address such limitations with deep learning. In [11], they employed word embedding technique for feature extraction and proposed a combined model of long short term memory (LSTM) and convolutional neural network (CNN) for detecting suicidal ideation in Reddit posts. [13] proposed a C-LSTM model for learning suicidal ideation connotations. The model utilizes CNN for extracting phrase representation and a LSTM for sentence representation. In the work of [14], they used a word embedding layer with Global Vectors for Word Representation (GloVe) as the input layer and a bidirectional LSTM (BiLSTM) to process the sequence of word vectors to predict the suicide risk from social media contents.

The increasing of suicide rates and the urgency to identify those who are at risk of suicide have motivated researchers to propose methods for suicidal ideation detection. Many methods have been proposed in the literature to detect suicidal ideation. However, current studies mainly focussing on detecting suicidal ideation from forum posts. There

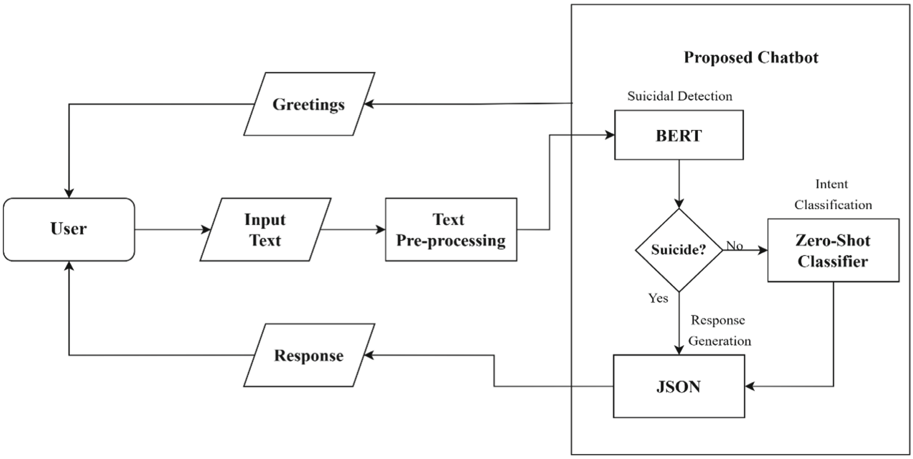


Fig. 1. A two-stage classification approach of our proposed chatbot

is limited research in using chatbot for suicidal ideation detection. To the best of our knowledge, this is the first work to investigate the use of chatbot in detecting suicidal ideation.

3 Proposed Method

We propose to combine a Bidirectional Encoder Representation from Transformer (BERT) with zero-shot learning in a two-stage classification chatbot system. The first stage is to determine whether individuals through their conversations with the chabot have suicidal ideation. The output of this stage will then be passed as input to the second stage. In this second stage, a zero-shot learning is applied for intent classification. Figure 1 shows our proposed chatbot for suicidal ideation detection.

3.1 First Stage: Suicidal Ideation Detection

In this stage, we applied a pre-trained language representation model based on BERT [15]. BERT is a multi-layered encoder with self-attention mechanism. Since BERT is bi-directionally trained, it can have a better representation of language context, which is one of the reasons why BERT is commonly applied for Natural Language Processing (NLP) tasks [16]. There are works that applied BERT for suicidal ideation [13, 17].

Our BERT model is trained on suicidal dataset obtained from Reddit Suicide Watch [18]. The data consists of 318797 records. We followed the standard approaches in text classification for text pre-processing. The posts are lowercased and tokenized, after removing stop words and punctuation marks. Stratified random sampling was applied to split the data into 4 partitions. Each partition consists of 15000 records for training and 10000 records for testing. The distribution of classes, "suicide" and "non-suicidal" for each partition of training and testing as shown in Table 1. The remaining 218797 records are used for out of sample testing.

Table 1. Distribution of data for each partition: (a) Training and (b) Testing

(a)

Data partition	Classes	
	Suicide	Non-suicidal
Partition 1	9000	6000
Partition 2	9292	5708
Partition 3	9324	5676
Partition 4	9232	5768

(b)

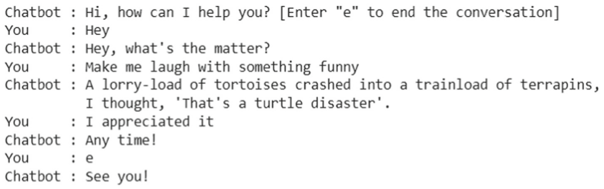
Data partition	Classes	
	Suicide	Non-suicidal
Partition 1	6100	3900
Partition 2	6215	3785
Partition 3	6162	3838
Partition 4	6192	3808

Table 2. Performance of BERT model in terms of accuracy, precision, recall and F1-score

Data partition	Performance measure			
	Accuracy	Precision	Recall	F1-score
Partition 1	0.94	0.92	0.92	0.92
Partition 2	0.94	0.92	0.91	0.92
Partition 3	0.94	0.94	0.90	0.92
Partition 4	0.94	0.91	0.93	0.92
Average	0.94	0.92	0.92	0.92

We used the Tensorflow library to train the BERT model. The model consists of an input layer, a BERT pre-processing layer, a BERT layer, a dropout layer with 0.4 dropout rate and a sigmoid layer running on NVIDIA Tesla T4 GPU.

Table 2 shows the performance of our model for each data partition. Accuracy, precision, recall and F1-score were used to measure the model's performance. Our BERT model achieved an average accuracy of 0.94 and 0.92 for precision, recall and F1-score. From the results, the BERT models trained on all the four partitions performed equally well. The model trained on the first partition was chosen as the final model since it has the same precision and recall rate, which is also the same as the average rate from the four partitions. This model was evaluated with the out of sample data and achieved an accuracy of 0.94, F1-score of 0.92. The results showed that our model is able to effectively detect suicidal ideation and was used to implement in the chatbot.



```
Chatbot : Hi, how can I help you? [Enter "e" to end the conversation]
You     : Hey
Chatbot : Hey, what's the matter?
You     : Make me laugh with something funny
Chatbot : A lorry-load of tortoises crashed into a trainload of terrapins,
          I thought, 'That's a turtle disaster'.
You     : I appreciated it
Chatbot : Any time!
You     : e
Chatbot : See you!
```

Fig. 2. Conversation with the chatbot

3.2 Second Stage: Intent Classification

One of the challenges in developing a chatbot is to recognise the ‘significance’ of a given text input in order for the chatbot to provide an appropriate response. From the first stage, if the user is not identified at risk of suicide, then it will proceed to the second stage for intent classification. Intent classification is generally performed to understand the intention behind the input text that the chatbot has received.

In this stage, we used the zero-shot model for intent classification. Unlike the traditional classification model, where it learns from a set of labelled training samples, zero-shot learns a classifier, where none of the classes are seen in the training set. Therefore, models trained with zero-shot learning are not required to retrain or fine-tune for specific task with the help of transfer learning from previously seen classes and auxiliary information. This makes zero-shot a suitable approach for intent classification.

We applied a pre-trained zero-shot classifier from the Transformer library to classify the intent of the input text. The aim of this stage is to enable the chatbot in providing suitable response based on the classified intent and to further engage with the user.

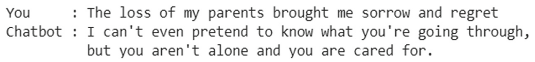
4 Implementation

We used a combination of contextual and keyword recognition-based methods to develop the chatbot. A list of responses is stored in the JavaScript Object Notation (JSON) file, where a response is retrieved based on the classified intent. The proposed two-stage classification approach (Sect. 3) is incorporated into the development of the proposed chatbot.

Figure 2 shows an example of a user having a conversation with the chatbot. The chatbot will first greet the user and prompt user for input. Once the user enters a text, the chatbot will pre-processed the text and pass to the BERT model to determine if the text has suicide risk. Since no suicide risk is detected, the input text will be passed on to the zero-shot classifier for intent classification. In this example, the user was requesting for a joke that will make him or her laugh.

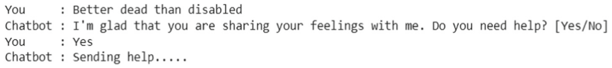
Besides detecting suicidal ideation, the chatbot is also able to detect user’s emotions. Figure 3 shows an example when a sad emotion is detected and the chatbot will comfort the user based on the predefined responses listed in the JSON file.

Figure 4 shows an example when suicidal ideation is detected. In this example, the BERT model classifies the input text as ‘suicide’. When suicide risk is detected, the chatbot will prompt the user if they need to get any help from the mental health professionals. If yes, then the chatbot can connect and alert the mental health professionals for



```
You     : The loss of my parents brought me sorrow and regret
Chatbot : I can't even pretend to know what you're going through,
          but you aren't alone and you are cared for.
```

Fig. 3. Sad emotion is detected



```
You     : Better dead than disabled
Chatbot : I'm glad that you are sharing your feelings with me. Do you need help? [Yes/No]
You     : Yes
Chatbot : Sending help.....
```

Fig. 4. Suicidal ideation is detected

needed care. Otherwise, the chatbot will continue to engage conversations with the user expressing concern about their welfare.

5 Conclusion

Current mental health professionals are insufficient to meet the demand to support those who may need treatment for mental health conditions. This could lead to more suicidal cases. Therefore, chatbots are suggested to be an alternative mental health support. In this paper, we proposed a two-stage classification chatbot for suicidal ideation detection. We have shown the viability of using chatbot in detecting suicidal ideation. We plan to extend our work to incorporate suicide intent scale to access the severity of the suicidal ideation through text conversation with the user. Assessing the level of the intent can help to provide appropriate responses.

References

1. World Health Organization, Comprehensive mental health action plan 2013–2030, 2021.
2. S. Park, Y. Lee, T. Youn, B. S. Kim, J. I. Park, H. Kim, H. C. Lee, J. P. Hong, "Association between level of suicide risk, characteristics of suicide attempts, and mental disorders among suicide attempters", BMC Public Health 18(1) (2018) 1–7. https://doi.org/10.1186/s12889-018-5387-8
3. B. Lew, K. Kõlves, D. Lester, W. S. Chen, N. B. Ibrahim, N. R. B. Khamal, F. Mustapha, C. M. H. Chan, N. Ibrahim, C. S. Siau, L. F. Chan, "Looking into recent suicide rates and trends in Malaysia: A comparative analysis", Frontiers in Psychiatry 12 (2022) 770252. https://doi.org/10.3389/fpsyt.2021.770252
4. P. Rytterström, S. M. Ovox, R. Wärdig, S. Hultsjö, "Impact of suicide on health professionals in psychiatric care mental healthcare professionals' perceptions of suicide during ongoing psychiatric care and its impacts on their continued care work", International Journal of Mental Health Nursing 29 (2020) 982-991. https://doi.org/10.1111/inm.12738
5. A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, J. B. Torous, "Chatbots and conversational agents in mental health: a review of the psychiatric landscape", The Canadian Journal of Psychiatry 64(7) (2019) 456-464. https://doi.org/10.1177/0706743719828977
6. A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, M. Househ, "An overview of the features of chatbots in mental health: A scoping review", International Journal of Medical Informatics 132 (2019) 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

7. M. Birjali, A. Beni-Hssane, M. Erritali, “Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks”, Procedia Computer Science, 113 (2017) 65-72. https://doi.org/10.1016/j.procs.2017.08.290
8. A. Mbarek, S. Jamoussi, A. Charfi, A. B. Hamadou, Suicidal profiles detection in Twitter, in: A. Bozzon, F. Domínguez Mayo, J. Filipe (Eds.), Proceedings of the 15th International Conference on Web Information Systems and Technologies, SCITEPRESS, Vienna, Austria, 2020, pp. 289–296. DOI: https://doi.org/10.5220/0008167602890296.
9. D. Ramírez-Cifuentes, A. Freire, R. Baeza-Yates, J. Puntí, P. Medina-Bravo, A. Velazquez, J. Gonfaus, J. Gonzàlez, “Detection of Suicidal Ideation on Social Media: Multimodal, Relational, and Behavioral Analysis”, Journal of Medical Internet Research, 20 (2020) e17758. https://doi.org/10.2196/17758
10. R.N. Grant, D. Kucher, A. M. León, J. F. Gemmell, D. S. Raicu, S. J. Fodeh, “Automatic extraction of informal topics from online suicidal ideation”, BMC Bioinformatics 19(211) (2018) 57-66. https://doi.org/10.1186/s12859-018-2197-z
11. A. L. Nobles, J. J. Glenn, K. Kowsari, B. Teachman, L. E. Barnes, Identification of imminent suicide risk among young adults using text messages, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, Montreal, Canada, 2018, pp. 1–11.
12. M. M. Tadesse, H. Lin, B. Xu, L. Yang, “Detection of suicide ideation in social media forums using deep learning”, Algorithms 13(1) (2020) 7. https://doi.org/10.3390/a13010007
13. R. Sawhney, P. Manchanda, P. Mathur, R. Shah, R. Singh, 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, ACL, Brussels, Belgium, 2018, pp. 167–175.
14. G. Coppersmith, R. Leary, P. Crutchley, A. Fine, “Natural language processing of social media as screening for suicide risk”, Biomedical informatics insights, 10 (2018) 1178222618792860. https://doi.org/10.1177/1178222618792860
15. J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Annual Conference of the American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Minneapolis, USA, 2019, pp. 4171–4186.
16. C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune BERT for text classification?, in: M. Sun, X. Huang, H. Ji, Z. Liu, Y. Liu (Eds.), Proceedings of the China National Conference on Chinese Computational Linguistics, Lecture Notes in Computer Science, vol. 11856, Springer, Cham, Kunming, China, 2019, pp. 194-206. https://doi.org/10.1007/978-3-030-32381-3_16
17. A. K. Ambalavanan, P. D. Jagtap, S, Adhya, M. Devarakonda, Using contextual representations for suicide risk assessment from Internet forums, in: K. Niederhoffer, K. Hollingshead, P. Resnik, R. Resnik, K. Loveys (Eds.), Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, ACL, Minneapolis, Minnesota, 2019, pp. 172–176. https://doi.org/10.18653/v1/W19-3022
18. S. Ji, C. P. Yu, S.-F. Fung, S. Pan, G. Long, “Supervised learning for suicidal ideation detection in online user content”, Complexity, 2018 (2018) 6157249. https://doi.org/10.1155/2018/6157249