人工智慧輔助社群媒體高風險訊息偵測系統

Artificial Intelligence Assisted Detection System for High Risk Messages in Social Media

Kun-Hua Lee [1], Kuo-Liang Ou[2] and Daw-Wei Wang[3*]

[1] Department of Educational Psychology and Counseling, National Tsing Hua University

[2]Institute of Learning Sciences and Technologies, National Tsing Hua University

[3] Department of Physics and Counseling Center, National Tsing Hua University

**E-mail: dwwang@phys.nthu.edu.tw*

In the last decade, social media has drastically changed the pattern of interpersonal interaction, especially for the young generation on campus. Literature indicated suicidal attempters could seek for helps to their peers, parents or relatives in different ways (especially in social media) before conducting suicide. Since counselors or mental professionals could not be aware of suicidal related messages immediately, early detection of these messages becomes very important. The purpose of this study is to establish an artificial intelligence (AI) assisted system to search for suicidal related high-risk messages in social media, providing alarming information to related security departments.

Our data was collected from a social media site, Dcard (https://www.dcard.tw), which is popular among college students. Total of 55,989 mood diaries from Dcard were collected by web crawlers. Each mood diary was assessed the level of mood intensity with an average mood score (total score divided by total number of words). Those scored lower than -1.4 would be grouped in A1 and those scored -1.2 to -1.4 were grouped in A2 in this study (Figure. 1). Three categories were encoded by mental professionals, including life events, special dates, and psychiatric service. We also classified eight types of paragraphs, including of positive emotion, positive reason, helplessness, hopelessness, physical arousal, suicide and depression, suicidal behaviors and negative emotion. Finally, four level of risk factors were identified: A (suicide/self-harm action), B (suicidal attempt), C (mild level) and 0 (none). Their number distribution can be found in Table. 1.

In our work, we first use natural language processing (NLP) technique to train an AI model to classify these labelled sentences, after certain data cleaning and segmentation. The total available sentences are 10533, including neutral sentences obtained randomly in other parts of the articles (see Table 2). Using Bi-LSTM (Bidirectional Long-Short Term Memory) as the fundamental algorithm, we have calculated the results for such a six-class model, obtaining an overall accuracy to be 63.92%. This is pretty good if considering so many classes. However, we also have found that classifying these negative sentences may not be trivial, while it could be very promising to distinguish negative sentences from positive/neutral ones(see Table 3). This information can be applied to the classification of crisis level of a mood diary.

Since the number of higher mood score messages is significantly more than the messages of lower mood score (Table1), the anomaly detection machine learning technique - OneClass Support Vector Machine (OCSVM) is applied to construct a model for predicting high-risk messages. Psychologists are employed to identify the types of each sentence and classify the training messages into the target crisis levels(Table 2). The result reveals that if all the training messages are classified into two classes, one class with crisis level A, B, and C, another class with crisis level 0 only. The precision, recall, and F1-score are all greater than 94% (see Table 4), which means the OCSVM model has an outstanding ability to recognize whether the messages are risky or not. Furthermore, The ratio of each type of sentences tagged

by experts is used as a feature for constructing another model for predicting. The results indicate that the accuracy is higher than 97%, meaning both the numbers and the ratio of each type of risk message could be utilized as a feather for classification.

However, if one class has crisis levels A and B, the other class has C and 0. The precision, recall, and F1-score are decreased to 70%, which means the OCSVM model cannot recognize the messages with higher risk. Therefore, the researchers may use the natural language processing (NLP) methods to construct another model to distinguish the high-risk message instead of considering the term-frequency only in the future.

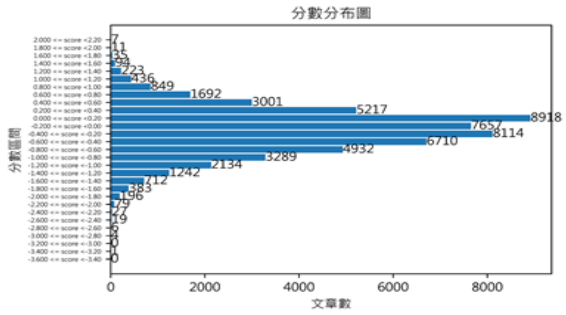

Figure. 1 Mood score distribution of all mood diaries in our dataset.

危機程度	A1	A2	總計	比例
A - 危險程度高	70	40	110	4.13%
B - 危險程度中	165	85	250	9.38%
C - 危險程度低	477	280	757	28.42%
0 - 無危機狀況	712	835	1547	58.07%
小計(篇數)	1424	1240	2664	100%

Table1. Crisis level distribution of datasets, A1 (scores < -1.4) and A2 (-1.4<scores<-1.2).

句子標籤	句數	比例	重新組合	句數
希望或幸福感	161	1.53%	正向	527
正向理由(認知)	366	3.47%		
中性(隨機選取)	1600	15.19%	中性	1600
無助(認知)	615	5.84%	無助無望	781
無望(認知)	166	1.58%		
生理反應	1175	11.16%	生理反應	1175
自殺行為	495	4.70%	自殺行為	4074
自殺與憂鬱	3579	33.98%		
負向情緒	2376	22.56%	負向情緒	2376
總計	10533	100.00%	總計	10533

Table 2. Classification of our sentence labeling and regrouping into 6 categories. The neutral sentences are collected by computer from other sentences without labelling.

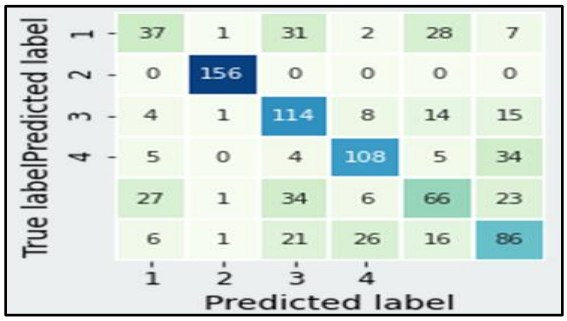

Table 3. Confusion matrix of our model in these six categories.

	precision	recall	f1-score	support
A,B,C	0.94	0.96	0.95	177
0	0.96	0.94	0.95	179
accuracy			0.95	
macro avg	0.95	0.95	0.95	356
weighted avg	0.95	0.95	0.95	356

Table 4. The prediction performance of our OCSVM model.