Neural Computing and Applications (2022) 34:10309–10319
https://doi.org/10.1007/s00521-021-06208-y

S. I.: EFFECTIVE AND EFFICIENT DEEP LEARNING

Suicidal ideation and mental disorder detection with attentive relation networks

Shaoxiong Ji[1] · Xue Li[2] · Zi Huang[2] · Erik Cambria[3]

Received: 4 June 2020 / Accepted: 8 June 2021 / Published online: 24 June 2021

Abstract
Mental health is a critical issue in modern society, and mental disorders could sometimes turn to suicidal ideation without effective treatment. Early detection of mental disorders and suicidal ideation from social content provides a potential way for effective social intervention. However, classifying suicidal ideation and other mental disorders is challenging as they share similar patterns in language usage and sentimental polarity. This paper enhances text representation with lexicon-based sentiment scores and latent topics and proposes using relation networks to detect suicidal ideation and mental disorders with related risk indicators. The relation module is further equipped with the attention mechanism to prioritize more critical relational features. Through experiments on three real-world datasets, our model outperforms most of its counterparts.

Keywords Suicidal ideation · Mental disorder · Attentive relation networks

1 Introduction

Mental health is a global issue, especially severe in most developed countries and many emerging markets. According to a report on mental health from the World Health Organization[1], 1 in 4 people worldwide suffers from mental disorders to some extent. Furthermore, 3 out of 4 people with severe mental disorders do not receive treatment, making the problem worse. Previous studies reveal that suicide risk usually has a connection to mental disorders [1]. Partly due to severe mental disorders, 900,000 persons commit suicide each year worldwide, making suicide the second most common cause of death among the young. Suicide attempters are also reported as suffering from mental disorders, with an investigation on the shift from mental health to suicidal ideation conducted by language and interactional measures [2]. The US National Alliance on Mental Illness reported that 46% of suicide victims had experienced mental health conditions[2]. According to the World Bank, at least 10 percent of the global population suffers from mental health issues[3].

With the advance of social network services, people begin to express their feelings in the forums and seek online support. Traditional ways of prevention include conversation-based consultation and psychological intervention. However, due to the scarcity and inequality of

S. Ji: The work was done while this author was at the University of Queensland.

✉ Erik Cambria
cambria@ntu.edu.sg

Shaoxiong Ji
shaoxiong.ji@aalto.fi

Xue Li
xueli@itee.uq.edu.au

Zi Huang
huang@itee.uq.edu.au

[1] Department of Computer Science, Aalto University, Espoo, Finland

[2] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia

[3] School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore

[1] Mental health action plan 2013–2020, available in http://www.who.int/mental_health/action_plan_2013/mhap_brochure.pdf?ua=1.

[2] NAMI report on Risk Of Suicide, available at https://www.nami.org/Learn-More/Mental-Health-Conditions/Related-Conditions/Suicide.

[3] Records updated on Apr 02, 2020. Available via https://www.worldbank.org/en/topic/mental-health.

public resources in health services [3], many victims could not get effective treatments even though some of them are suffering from severe mental disorders.

Transferring from mental disorders to suicidal ideation and the final suicide action is a long-term process. Gilat et al. [4] scaled suicide risks into four levels, i.e., non-suicidal, suicidal thoughts or wishes, suicidal intentions, and suicidal act or plan. Before suicidal ideation, victims may suffer from different kinds of other mental disorders. According to meta-analyses, underlying mental disorders can lead to suicide, mainly in high-income countries with a figure of 90%. According to meta-analyses by Hannah Ritchie and Max Roser[4], underlying mental disorders can lead to suicide. This is especially prevalent in high-income countries, with 90% of people who encounter mental health issues have suicidal thoughts. The social networking service has become one of the most valuable tools to provide support and feedback for people with mental health issues [5]. To provide effective suicide early prevention given limited support resources, it is necessary to triage the risk levels automatically and provide conversational support accordingly to relieve victims' issues [6]. Our motivation is to use deep learning techniques to enable early detection and identify people's risk levels, which can help the social workers or experts to have a prior understanding of people's situation when trying to relieve their mental health issues. The automatic detection technique can be applied to mental health monitoring and help to facilitate online support. We conduct suicidal ideation and mental health detection, aiming to distinguish early-stage mental disorders and severe suicidality automatically. This classification of suicide risk and mental status could help social workers prioritize and allocate resources to people with different needs and situations according to their severity. Thus, effective prevention measures can be taken to stop mental health discourse transitions to suicidal thoughts.

Suicide and mental health issues could be categorized as different levels, taken as a multi-class classification problem. There are many types of mental disorders according to two main diagnostic schemes for identifying mental disorders, i.e., Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[5] and Chapter V Mental and Behavioral Disorders of International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)[6]. Recent works using deep neural networks have revolutionized this field of text classification. However, classifying mental health and suicidal ideation is a more specific task that requires focusing on potential victims' language usage. Suicidal ideation and mental disorders (e.g., depression, anxiety, and bipolar) in online social content share quite similar patterns, including the language usage [7], topic distribution [8], and sentiment polarity [9] . Affective computing research [10] is employed for assigning emotions to suicide notes [11], where most of the notes contain many negative expressions. Topics including job stress, family issues, and personal crisis are quite common among those posts [8]. Thus, classifying suicidal ideation and other mental health issues requires attention to understand the subtle differences among those characteristics. Noticing that people's posting showing feelings or expressing suffering contains their sentiment to some extent, we propose to capture this valuable and vital information to learn richer sentence representation and better encode risk actions and people's mental or social state.

This paper investigates deep learning-based models for text classification on some existing and self-collected datasets from social networking websites. It proposes an enhanced relation network (RN) to provide a more accurate classification of suicide risk levels and suicidal ideation vs. other mental disorders such as depression and anxiety. Relation networks [12] are firstly proposed to visual reasoning, while the undefined relationship between texts and sentiment lexicons or relationship between texts and topic distributions remains unexplored. In this paper, we firstly migrate the principle of visual relational reasoning to relate extracted informative features and hidden text representation and develop a novel attentive encoding model to capture the relation between suicide indicators such as sentiments and event topics, and text mentions. This novel relational encoding model can reason over the risk indicators and sentence embeddings and learn richer representations.

Our contributions could be summarized as:

- This paper focuses on identifying suicidal ideation and different kinds of mental disorders for early warning. Specifically, we consider both user-level and post-level detection.
- To improve risk identification performance, we propose an attentive RN model with text representation and two sets of risk indicators encoded, i.e., lexicon-based sentimental state and latent topics within posts.
- Experiments on public datasets and our collected dataset show that our proposed method can improve the predictive performance.

This paper is organized as follows. Related work on mental disorders, suicidal ideation, text classification, and relational reasoning are reviewed in Sect. 2. In Sect. 3, we introduce the proposed method that introduces sentimental lexicon and topic model into relational encoding with

[4] Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/mental-health.

[5] http://psychiatry.org/psychiatrists/practice/dsm.

[6] http://apps.who.int/classifications/icd10/browse/2016/en#/V.

attentive RN. Datasets are introduced in Sect. 4, together with a simple exploratory analysis. Experimental settings and results are presented in Sect. 5. In Sect. 6, we conclude and have a brief outlook for future work.

2 Related work

This paper focuses on enabling effective relational text encoding to classify suicidal ideation and mental disorders. Related works include research on mental disorders, suicidal ideation, text classification techniques, and relational reasoning.

Mental health issues and suicidal ideation [6] have been studied including the clinical interaction [13], classifying self-report screening questionnaire [14], and detection from the data mining perspective [15, 16]. With the popular social text analysis and natural language processing techniques, more and more research turns to investigate the mental health discourse [17], discover self-harm content [18], and detect social network mental disorders [19], depression [20] and suicidal ideation [8] in social media. Li et al. [21] detected the changes of online users in the mental health communities of social media. Cao et al. [22] proposed to use a personal knowledge graph to improve the detection performance. Affective information is widely used for mental health and suicidality detection. Nguyen et al. [23] proposed a thorough affective analysis and a content analysis between depression and control communities. Ren et al. [24] proposed an accumulated emotion model to classify suicide blog stream. Chen et al. [9] measured emotions to identify depression on Twitter. We recommend readers to check out the recent review article [6] about suicidal ideation detection with a detailed introduction and summary.

Suicidal ideation and mental disorder detection are technically formulated as a text classification, which has experienced a rapid development with the development of deep neural networks [25]. Kim [26] proposed convolutional neural networks for sentence classification. To capture long-term dependencies in sentences, the long short-term memory (LSTM) [27] was applied. Lai et al. [28] proposed recurrent convolutional neural networks combining two popular neural network architectures for text classification. Li et al. [29] combined reinforcement learning, generative adversarial networks, and recurrent neural networks for text categorization. Zhao et al. [30] explored the use of capsule networks for challenging natural language processing applications. The attention mechanism [31] is also widely used in text classification. For example, Lin et al. [32] proposed self-attention to learn structured sentence embedding, and Ma et al. [33] proposed an attentive LSTM for aspect-based sentiment analysis.

Those informative affective cues and neural advances inspire this paper to develop a deep model to learn hidden text representation and encode extracted features such as emotion information [9] and topic descriptions [8], and external resources like domain-specific lexicons [34] in a hybrid manner. This paper incorporates relational reasoning with relation networks (RNs) to inject and fuse that multi-channel auxiliary information and rich textual representations. RNs are initially utilized for scene object discovery by exploiting relations among objects [12], and further introduced to relational reasoning for visual question answering by calculating the relation score of the feature maps of object pairs and question representation [35]. As for our application scenario of suicidal ideation detection, it is critical to understand the relation between suicidality and risk indicators such as an individual's sentiment and life events. This is also the novelty of this paper, i.e., to reason over the relation between risk indicators between text mentions using relation networks.

3 Methods

3.1 Problem definition

Detecting suicidal ideation and mental disorders in social content is technically a domain-specific task of text classification and social data analytics [36]. In our paper, we conduct a fine-grained suicide risk assessment and classification of multiple mental health issues, which are naturally regarded as multi-class classification. For fine-grained suicide risk, the risk levels include none, low, moderate, and severe risk, while for mental health classification, specific mental disorders are depression, anxiety, and bipolar. Moreover, there are two subtasks for specific social content settings, i.e., post-level classification and user-level classification. The former one takes single post p as input, while the latter one detects the suicide attempter with multiple posts $P = \{p_1, p_2, \ldots, p_n\}$.

3.2 Model architecture

The proposed model consists of two steps, i.e., post representation and relational encoding module as illustrated in Fig. 1. The post representation includes two parts of extraction of risk-related state indicators and LSTM text encoder. The relation module, as shown in the dashed box of Fig. 1 utilizes a vanilla RN for reasoning on the connection between state indicators and user's posts and the attention mechanism for prioritizing more important relation scores of text encoding.

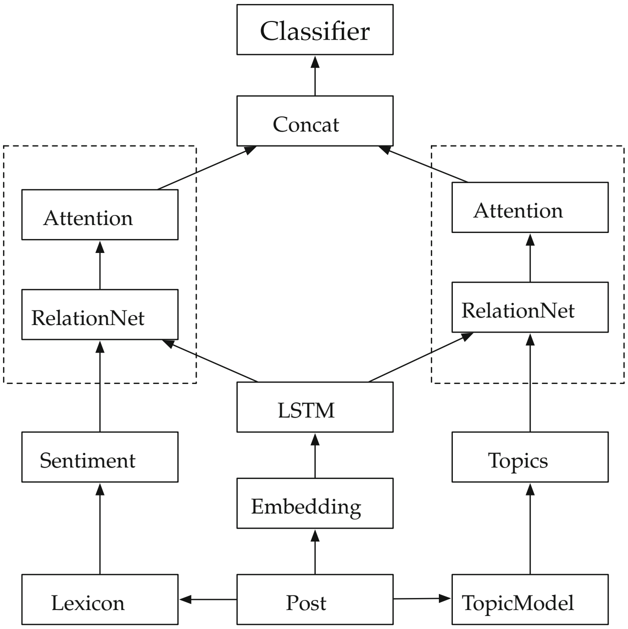


Fig. 1 The architecture of the proposed model

3.3 Text encoding and risk indicators

User post sequence is embedded into word vectors of $p = \{w_1, w_2, \ldots, w_n\} \in \mathbb{R}^{l \times d}$, where l is the length of posts and d is the dimension of word embeddings. We apply bidirectional LSTM in Eq. 1 for text encoding to capture the adjacent dependency of words.

$$\begin{aligned} \overrightarrow{h_t} &= \overrightarrow{\text{LSTMcell}}\left(w_t, \overrightarrow{h_{t-1}}\right) \\ \overleftarrow{h_t} &= \overleftarrow{\text{LSTMcell}}\left(w_t, \overleftarrow{h_{t+1}}\right) \end{aligned} \tag{1}$$

The hidden state is obtained by concatenating each direction as $h_t = \text{concat}(\overrightarrow{h_t}, \overleftarrow{h_t})$, where $h_t \in \mathbb{R}^{l \times 2n}$ given n as the number of hidden units.

Sentimental information plays an important role when people are expressing their sufferings and feelings in online social networks. To measure the sentiment, we take sentiment lexicons as additional information. Specifically, domain-specific sentiment lexicons [34] from communities in Reddit are used. Seed words induce the sentiment lexicons with domain-specific word embedding and a label propagation framework. For the details of building domain-specific sentiment lexicons, we recommend readers refer to the original paper. The extracted sentiment information of post denoted as $s \in \mathbb{R}^l$ acts as a state indicator representing post creators' internal sentimental state. Correspondingly, extrinsic indicators such as people's event topics reveal another dimension as the risk indicator. To capture external factors of suicidal ideation or mental disorders, we introduce a topic model to learn unsupervised topical features. Specifically, Latent Dirichlet Allocation (LDA) [37] is applied to extract latent topics in social posts to represent people's sufferings such as life events, social exposure, and other experience in the real world. The probability score vectors of posts belonging to all extracted topics are represented as $v \in \mathbb{R}^m$, where m is the number of topics.

3.4 Relation network with attention

RN [35] is a neural module for relational reasoning. It is originally proposed to capture the relation between objects. Given objects of $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$ and functions of f_ϕ and g_θ, a RN is defined in Eq. 2. The output of g_θ is called the learned "relation", while the f_ϕ function acts as the classifier.

$$\text{RN}(\mathcal{O}) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right) \tag{2}$$

We aim to encode risk factors of suicidal ideation and mental disorders into textual representation. Thus, we take text encoding and state indicators as the input of RNs to calculate relation scores between each token in posts and state indicators modeled by sentiment and topic features. The attention mechanism is further incorporated with the relation module by assigning attention weights to the learned relations. The idea of attentive RN is shown in Fig. 2. The text representation is encoded by an LSTM network, which captures the sequential independence. The encoded text representation is then concatenated with the representations of state indicators. Here, we consider two indicators of sentiment and topic features, with the expanded representations by repeating the extracted vectors denoted as $S = [s, s, \ldots, s] \in \mathbb{R}^{l \times l}$ and $V = [v, v, \ldots, v] \in \mathbb{R}^{l \times m}$ respectively. Then, they are inputted into RNs to calculate relation vector $r_i \in \mathbb{R}^k$, where k is the dimension of hidden representation, with a multiple layer perception (MLP) as in Eq. 3 for the sentiment indicator.

$$r_i = \text{MLP}(h_i, s_i) \tag{3}$$

The attention is calculated as follows:

$$\alpha = \text{softmax}\left([r_1, r_2, \ldots, r_l]W^T + b\right), \tag{4}$$

where $W \in \mathbb{R}^{1 \times k}$, $b \in \mathbb{R}^l$ and $\alpha \in \mathbb{R}^l$. By element-wise product, the attentive representation of learnt relations can be calculated as

$$\tilde{r} = \alpha \otimes [r_1, r_2, \ldots, r_l] \tag{5}$$

where $\tilde{r} \in \mathbb{R}^{l \times k}$. By applying element-wise sum over $\tilde{r}$, we get the final attentive relational representation.

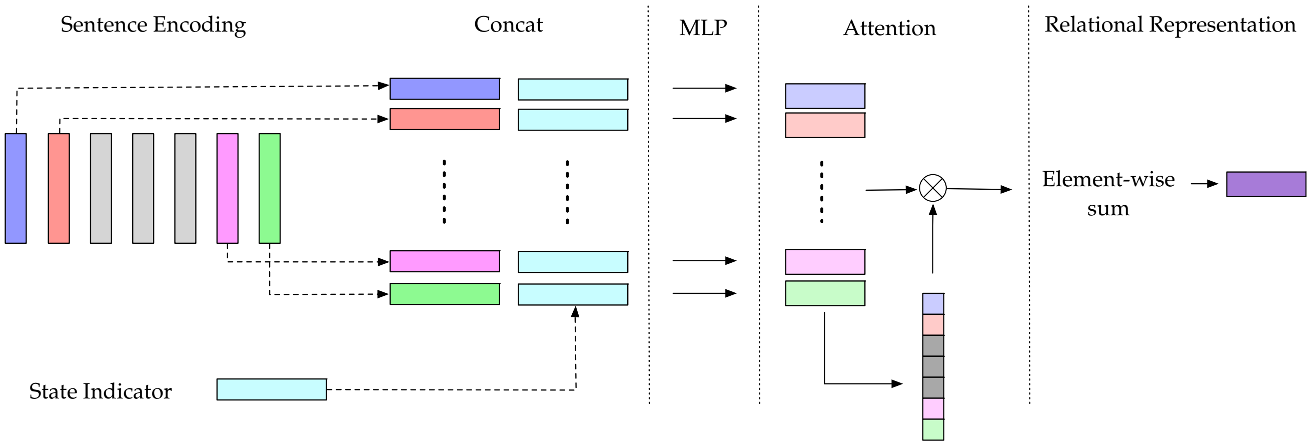


Fig. 2 Relation network with attention mechanism

3.5 Classification

The last step is to use the learned representation, which contains sequential information and risk indicators for classification. Specifically, we concatenate relational representations of $e = [\tilde{r}_s, \tilde{r}_v]$ from two channels as shown in Fig. 1, and use the fully connected layer with nonlinear activation function of $f(\cdot)$ to produce the logits for prediction as follows.

$$\begin{aligned} l &= \mathrm{f}(W_l e + b_l) \\ \mathcal{P} &= \mathrm{softmax}(W_o l + b_o) \end{aligned} \tag{6}$$

where $W_l \in \mathbb{R}^{d_l \times d_e}, b_l \in \mathbb{R}^{d_l}, W_o \in \mathbb{R}^{c \times d_l}, b_o \in \mathbb{R}^c, \mathcal{P} \in \mathbb{R}^c$ For multi-class classification, the predicted label is produced by

$$\hat{y} = \underset{i}{\mathrm{argmax}}(\mathcal{P}_i) \tag{7}$$

3.6 Training

Our proposed model has two training phases, i.e., training the LDA topic model and the classification model. For the topic model, LDA assumes a generative process of documents as random mixtures over latent topics. A topic can be inferred as a distribution over the words, where the Bayesian inference is used for learning various distributions. In practice, we use the Gensim library[7] to build the topic model during implementation.

For the ultimate target of suicide ideation and mental health detection, we use the cross-entropy loss with L2 regularization, denoted as:

$$L = -\frac{1}{\sum_{s=1}^{N} c(s)} \sum_{i=1}^{N} \sum_{j=1}^{c(i)} \log \mathcal{P}_{i,j}\left[y_{i,j}\right] + \lambda \|\theta\|_2, \tag{8}$$

where $c(s)$ is the set of labels, θ represents all the trainable parameters, and λ is the regularization coefficient or the so-called weight decay rate. We apply the Adam algorithm [38] to optimize the objective function.

4 Data

We use data from two popular social networking websites, i.e., Reddit and Twitter, with three datasets derived. Two of them are from Reddit with one public dataset and one firstly collected in this paper. People's posts from an active subreddit for online support in Reddit, called "SuicideWatch"(SW)[8], are intensively used in these two datasets. The last one is collected from Twitter by combining several existing data sources. These datasets cover suicide and other mental health issues, with specific categories reported in ICD-10 as listed in Table 1.

4.1 UMD reddit suicidality dataset

The UMD Reddit Suicidality Dataset [39] was collect from anonymous discussion forums in https://www.reddit.com. It contains posts of 620 users in the training set and 245 users in the testing set sampled from 11,128 users in the subreddit "SuicideWatch" and 11,129 users in other subreddits. It is annotated by crowdsourcing workers and human experts via a crowdsourced platform, referring to the original paper for annotations' details. The suicide risk is scaled to four levels, i.e., no risk(a), low(b), moderate(c), and severe risk(d). It also provides coarse labels where no risk and low risk are given the label of 0, moderated, and severe risk are labeled as 1, together with the control group as the label of −1.

[7] http://radimrehurek.com/gensim.

[8] http://reddit.com/r/SuicideWatch.

Table 1 Description of mental disorders in ICD-10

Category	Descriptions mentioned in ICD-10
Suicide	Intentional self-harm, suicidal ideation (tendencies)
Depression	In typical mild, moderate, or severe depressive episodes, the patient suffers from lowering of mood, reduction of energy, and decrease in activity
Anxiety	Phobic anxiety and other anxiety disorders
Bipolar	A disorder characterized by two or more episodes in which the patient's mood and activity levels are significantly disturbed
PTSD	Arises as a delayed or protracted response to a stressful event or situation (of either brief or long duration) of an exceptionally threatening or catastrophic nature

This dataset was released as the CLPsych 2018 Shared Task [39], and then a new version of it acted as the CLPsych 2019 Shared Task [40]. This paper uses a dataset derived from the UMD dataset in user-level four categories of suicide risk. The statistical information of the dataset is illustrated in Table 2. Besides, we include control users (labeled as "None") into this annotation set.

We use the transformed labels from the raw label according to the original description of this dataset. Specifically, raw labels of "c" or "d" are transformed into 1, raw labels of "a" or "b" are transformed into 0, and the label of a control user is -1 by definition. We split the whole dataset into training, validation, and testing sets as listed in Table 3.

4.2 Reddit SWMH dataset

As severe mental health issues are very likely to lead to suicidal ideation, we also collect another dataset from some mental health-related subreddits in https://www.redditinc.com/. to further the study of mental disorders and suicidal ideation. We name this dataset as Reddit SuicideWatch and Mental Health Collection, or SWMH for short, where discussions comprise suicide-related intention and mental disorders like depression, anxiety, and bipolar. We use the Reddit official API[9] and develop a web spider to collect the targeted forums. This collection contains a total of 54,412 posts. Specific subreddits are listed in Table 4, as well as the number and the percentage of posts collected in the train-val-test split.

In those communities or so-called subreddits, people discussed their own or their relative's mental disorders and sought advice or help. We perform experimental analysis on this dataset to identify discussions about suicidality and mental disorders.

4.3 Twitter datasets collection

The third dataset is a collection of different subsets from Twitter with an overlapping check. Sampled instances from two datasets consist of most samples of this dataset. First, 594 instances of tweets containing suicidal ideation are from Ji et al. [8], with additional 606 tweets manually collected by this work. Second, the same number of depression and post-traumatic stress disorder (PTSD) posts are sampled from CLPsych 2015 shared task dataset [41]. This dataset is available upon request[10]. Last, the control group where Twitter users are not identified as having a mental condition or suicidal ideation is comprised by sampling regular tweets from previously mentioned datasets [8, 41]. Finally, this Twitter dataset collection contains 4800 tweets with four classes of suicidality, depression, PTSD, and control.

4.4 Linguistic clues and emotion polarity

We have a brief exploratory analysis of the data. Some selected linguistic, statistical information of UMD dataset extracted by Linguistic Inquiry and Word Count software (LIWC)[11] is shown in Table 5. The risk of suicide increases among labels of -1, 0, and 1. The linguistic inquiry results show that negative emotion, anxiety, and sadness are expressed more in posts with high-level suicide risk. The same trend exists in family issues, death-related mentions, and swear words. Naturally, positive emotions are less present in posts with high suicide risk.

5 Experiments

To evaluate our proposed model's performance, we compare it with several text classification models on three real-world datasets. Baselines and empirical settings are introduced, and results are reported and discussed in this section.

5.1 Baseline and settings

We compared five popular classification models with our proposed method. These baseline models are described as follow:

- *fastText* [42]: an efficient text classification model with bag of words sentence representation and a linear classifier.

[9] http://reddit.com/dev/api.

[10] Request for data access via http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/.

[11] http://liwc.wpengine.com.

Table 2 Statistical information of UMD reddit suicidality dataset

Annotation	Numbers	% of a/b/c/d/ levels
Crowd	621	26%/10%/24%/40%
Expert	245	29%/9%/25%/37%

Table 3 Statistical information of UMD dataset with train/validation/test split

Label	#/% of train	#/% of valid.	#/% of test
−1	495/49.8489%	126/50.6024%	245/50.000%
0	188/31.2185%	89/35.7430%	86/17.551%
1	310/18.9325%	34/13.6546%	159/32.449%

Table 4 Statistical information of SuicideWatch and mental health-related subreddits, i.e., SWMH dataset

Subreddit	#/% of train	#/% of valid.	#/% of test
Depression	11,940/34.29	3032/34.83	3774/34.68
SuicideWatch	6550/18.81	1614/18.54	2018/18.54
Anxiety	6136/17.62	1508/17.32	1911/17.56
Offmychest	5265/15.12	1332/15.30	1687/15.50
Bipolar	4932/14.16	1220/14.01	1493/13.72

Table 5 Selected linguistic statistical information of UMD dataset extracted by LIWC

Linguistic clues	Label −1	Label 0	Label 1
Positive emotion	3.30	3.12	2.96
Negative emotion	1.56	2.30	2.74
Anxiety	0.17	0.33	0.41
Sadness	0.28	0.50	0.68
Family	0.29	0.39	0.47
Friend	0.43	0.56	0.54
Work	2.54	1.92	1.80
Money	1.13	0.71	0.61
Death	0.22	0.29	0.36
Swear words	0.23	0.33	0.40

- *CNN* [26]: it applies convolutional neural networks over the word embedding of sentence to produce feature maps and then uses max-pooling over the features.
- *LSTM* [27]: it takes sequential word vectors as input to the recurrent LSTM cells applies pooling over the output to obtain final representation. By combining the forward and backward direction, it becomes bidirectional LSTM (BiLSTM).
- *RCNN* [28]: this model at first applies LSTM model [27] to capture sequential information, and then applies CNN [26] to further extract features. It has a bidirectional version using BiLSTM.
- *SSA* [32]: it proposed a structured self-attention mechanism with multiple hops by introducing a 2D matrix for embedding representation. The self-attention is applied to the sequential hidden states of the LSTM network.

All the baseline models and our proposed method are implemented by PyTorch[12] and run in a single GPU (Nvidia GeForce GTX 1080 Ti). We train the models for 50 epochs by default, setting the batch size to be 128 and 16 according to the size of datasets. Specifically, the UMD dataset's batch size is 16, and for SWMH and Twitter data collection, the batch size is 128. We use a pretrained GloVe [43] word representation with either static or dynamic embedding utilized for the word embedding. Our proposed method enumerates all the 250 subreddit lexicons of Reddit and the number of topics from 5 to 20. We select the best validation performance in multiple trials and report the testing performance as experimental results.

Our RN-based model uses an additional attentive relation network, and SSA applies a structured attention network. Thus, SSA and RN have more trainable parameters compared with vanilla CNN and LSTM. Our model size is on par with the SSA model. In our experimental setting, we use 100D word embeddings and set the hidden dimension of the LSTM unit to 300. As our model uses another MLP to calculate the relation score, it consumes more parameters, around 30K parameters, than the SSA model. However, it is worth sacrificing the model size to achieve higher predictive performance for an essential mission of saving lives.

The goal of automatic detection is to produce effective diagnoses (i.e., true positive) and decrease the incorrect diagnoses (i.e., false positive) to avoid patients' stress and anxiety caused by false detection. Thus, during the evaluation process, we report both prediction accuracy and the weighted average F-score metric. For unbalanced datasets, we apply weight penalty to the objective function and report the weighted average results.

5.2 Results

We evaluate the experimental performance on three datasets collected from Reddit and Twitter. For the UMD

[12] http://pytorch.org.

Suicidality dataset and the SWMH dataset, the reported results are weighted average.

5.2.1 UMD suicidality

We firstly implement our method and baselines on the UMD suicidality dataset for user-level classification. When processing a set of posts from users, all posts of users are concatenated as user-level representation. Concatenated posts are truncated or padded with zeros to ensure an identical dimensionality. The results of four metrics, including accuracy, precision, recall, and F1-score, are very close for RN and BiLSTM (Table 6). The BiLSTM model gains the highest accuracy of 56.94%, and our model follows at the second place of 56.73%. Our model equipped with relational encoding, however, has a higher F1 score than all the baselines. Noticing these very close results, we then further analyze each class's results in Section 5.3.

5.2.2 Reddit SWMH

Then, we perform experiments on the Reddit SWMH dataset, which contains both suicidal ideation and mental health issues, to study our model's predictive performance. It is a larger dataset with more instances when compared with the UMD dataset. Experiments on this dataset show RN's relational encoding capacity for mental health-related texts with similar characteristics. As shown in Table 7, our model beats all baseline models in terms of all four metrics.

5.2.3 Twitter collection

Lastly, we conduct experiments on the Twitter dataset with similar settings to previous experiments. Unlike posts in Reddit, tweets in this dataset are short sequences due to the tweet's length limit of 280 characters. The results of all baseline methods and our proposed method are shown in Table 8. Among these competitive methods, our model gains the best performance on these four metrics, with 1.77% and 1.82% improvement than the second-best BiLSTM model in terms of accuracy and F1-score, respectively. Our proposed method introduces auxiliary information of lexicon-based sentiment and topics learned from the corpus and utilizes RNs to model the interaction between LSTM-based text encodings and risk indicators. Richer auxiliary information and efficient relational encoding help our model boost performance in short tweet classification.

5.3 Performance on each class

This section studies the performance of each class of the UMD dataset. We select two baselines with better performance for comparison as shown in Table 9. The proposed RN-based model is flawed in predicting posts without suicidality but good at predicting posts with high suicide risk. Unfortunately, all these three models have an inferior capacity for predicting posts with low suicide risk. In the UMD dataset with a small volume of instances, these models tend to predict posts as classes with more instances, even though we apply penalty on the objective function.

5.4 Ablation study

We then conduct an ablation study to explore several variants and compare their performance. We compare our complete framework with three different settings of injecting risk indicators. The BiLSTM+concat model concatenates the final hidden state with sentiment and topic features. The BiLSTM+RN+sentiment and BiLSTM+RN+topic use single-channel relation network. The results on the Twitter collection are shown in Table 10. Relation networks are generally better than BiLSTM with simple feature concatenation, where the performance of the latter model decreased compared to

Table 6 Comparison of different models on UMD dataset for user-level classification, where precision, recall, and F1 score are weighted average

Model	Accuracy	Precision	Recall	F1
FastText	0.5327	0.5300	0.5327	0.5202
CNN	0.5531	0.4498	0.5531	0.4935
LSTM	0.5612	0.4625	0.5612	0.5071
BiLSTM	**0.5694**	0.5029	**0.5694**	0.5233
RCNN	0.5592	0.4953	0.5592	0.5111
SSA	0.5633	0.4711	0.5633	0.4839
RN	0.5673	**0.5405**	0.5673	**0.5453**

Table 7 Comparison of different models on Reddit SWMH collection, where precision, recall, and F1 score are weighted average

Model	Accuracy	Precision	Recall	F1
FastText	0.5722	0.5760	0.5722	0.5721
CNN	0.5657	0.5925	0.5657	0.5556
LSTM	0.5934	0.6032	0.5934	0.5917
BiLSTM	0.6196	0.6204	0.6196	0.6190
RCNN	0.6096	0.6161	0.6096	0.6063
SSA	0.6214	0.6249	0.6214	0.6226
RN	**0.6474**	**0.6510**	**0.6474**	**0.6478**

Table 8 Performance comparison on Twitter dataset, where precision, recall, and F1 score are weighted average

Model	Accuracy	Precision	Recall	F1
FastText	0.7927	0.7924	0.7927	0.7918
CNN	0.7885	0.7896	0.7885	0.7887
LSTM	0.8021	0.8094	0.8021	0.8039
BiLSTM	0.8208	0.8207	0.8208	0.8195
RCNN	0.8094	0.8089	0.8094	0.8090
SSA	0.8156	0.8149	0.8156	0.8152
RN	**0.8385**	**0.8381**	**0.8385**	**0.8377**

Table 9 Performance on each class of UMD suicidality dataset

Label	Metrics	BiLSTM	SSA	RN
−1	Precision	0.62	0.57	0.69
	Recall	0.77	**0.92**	0.70
	F1-score	0.69	0.70	0.69
1	Precision	0.51	0.57	0.48
	Recall	0.55	0.31	**0.62**
	F1-score	0.53	0.41	0.54
0	Precision	0.15	0.00	0.24
	Recall	0.02	0.00	0.09
	F1-score	0.04	0.00	0.13

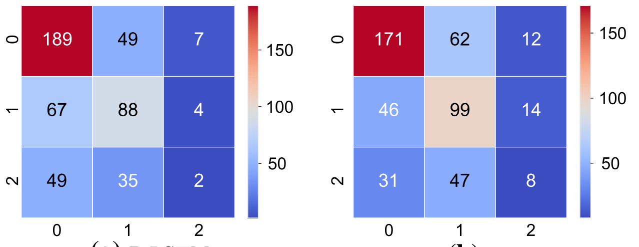
vanilla BiLSTM. The BiLSTM+RN+sentiment model is better than vanilla BiLSTM, while the BiLSTM+RN+topic is slightly worse. However, two channels of sentiment and topic couple well, achieving the best performance. This study shows the effectiveness of the proposed model.

5.5 Error analysis and limitations

This section conducts error analysis, taking UMD dataset as an example. As mentioned before in the last section, most methods suffer from poor performance in predicting low-risk posts. Figure 3 shows the heat maps of the confusion matrix of BiLSTM and our RN-based model, where axis 0, 1, and 2 represent labels of −1, 1, and 0. These two methods tend to predict more instances as none or high risk. Furthermore, our proposed method has a slightly better result than its counterpart. We also notice that our proposed model can achieve a higher accuracy of 59.18% on the UMD dataset. However, it fails in predicting low-risk suicidal ideation, with a similar performance to other baselines.*Limitations and Future Work.* We use sentiment lexicons and a topic model to exact sentiment- and topic-related risk indicators for relational text encoding. This preprocessing procedure can cause error propagation. Sentiment varies in different social communities. Using existing lexicons in popular communities may have a limitation. In future work, we will consider building lexicons from mental health-related communities. Machine learning-based based automatic detection systems cannot be treated as professional medical diagnoses. However, they can empower professional practitioners and help them identify potential victims from many social posts. The reasons for mental disorders and suicidal ideation are complex. Our future work will explore mental and suicidal factors and more effective relational reasoning to boost predictive performance.

Table 10 Performance of different variants of risk indicator injection

Model	Accuracy	Precision	Recall	F1
BiLSTM	0.8208	0.8207	0.8208	0.8195
BiLSTM+concat	0.8167	0.8262	0.8167	0.8190
BiLSTM+RN+sentiment	0.8240	0.8246	0.8240	0.8239
BiLSTM+RN+topic	0.8198	0.8177	0.8198	0.8183
BiLSTM+RN+sent.+topic	**0.8385**	**0.8381**	**0.8385**	**0.8377**



(a) BiLSTM **(b)** RN

Fig. 3 Confusion matrix on UMD dataset

6 Conclusion

Text classification on mental disorders might not be treated as a medical diagnosis of professional practitioners. It can act as a computer-aided system to automatically provide early warnings of online social users at risk and notify social workers to provide early intervention. It can also help the social workers and volunteers to identify the type of mental disorders, relieve online users' mental health issues through conversations, and suggest proper consultations or treatments. This paper attempts to encode text by integrating suicidal ideation with sentimental indicators and life event-related topical indicators and proposes RNs with an attention mechanism for relational encoding.

Experiments show the effectiveness of our proposed model. We argue that it is a significant step to combine canonical feature extraction with RNs for reasoning.

Data availability The authors would like to thank Philip Resnik for providing the UMD Reddit Suicidality Dataset (available via http://users.umiacs.umd.edu/~resnik/umd_reddit_suicidality_dataset.html.) and Mark Dredze for providing the dataset (available via http://www.cs.jhu.edu/~mdredze/datasets/clpsych_shared_task_2015/.) in the CLPsych 2015 shared task. This research is supported by the Australian Research Council (ARC) Linkage Project (LP150100671).

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. Windfuhr K, Kapur N (2011) Suicide and mental illness: a clinical review of 15 years findings from the uk national confidential inquiry into suicide. Br Med Bull 100(1):101–121
2. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M (2016) Discovering shifts to suicidal ideation from mental health content in social media. In CHI, pages 2098–2110. ACM
3. Jacob KS, Patel V (2014) Classification of mental disorders: a global mental health perspective. Lancet 383(9926):1433–1435
4. Gilat I, Tobin Y, Shahar G (2011) Offering support to suicidal individuals in an online support group. Arch Suicide Res 15(3):195–206
5. Shepherd A, Sanders C, Doyle M, Shaw J (2015) Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. BMC Psychiat 15(1):29
6. Ji S, Pan S, Li X, Cambria E, Long G, Huang Z (2021) Suicidal ideation detection: a review of machine learning methods and applications. IEEE Trans Comput Soc Syst 8(1):214–226
7. Coppersmith G, Leary R, Whyne E, Wood T (2015) Quantifying suicidal ideation via language usage on social media. In Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM
8. Ji S, Yu CP, Fung SF, Pan S, Long G (2018) Supervised learning for suicidal ideation detection in online user content. Complexity. https://doi.org/10.1155/2018/6157249
9. Chen X, Sykora MD, Jackson TW, Elayan S (2018) What about mood swings: Identifying depression on twitter with temporal measures of emotions. In The Web Conference, pages 1653–1660
10. Wang Z, Ho S, Cambria E (2020) A review of emotion sensing: categorization models and algorithms. Multimed Tools Appl 79:35553–35582
11. McCart JA, Finch DK, Jarman J, Hickling E, Lind JD, Richardson MR, Berndt DJ, Luther SL (2012) Using ensemble models to classify the sentiment expressed in suicide notes. Biomedical informatics insights, 5: BII–S8931
12. Raposo D, Santoro A, Barrett D, Pascanu R, Lillicrap T, Battaglia P (2017) Discovering objects and their relations from entangled scene representations. arXiv preprint arXiv:1702.05068
13. Venek V, Scherer S, Morency LP, Pestian J (2017) Adolescent suicidal risk assessment in clinician-patient interaction. IEEE Trans Aff Comput 8(2):204–215
14. Delgado-Gomez D, Blasco-Fontecilla H, Sukno F, Ramos-Plasencia MS, Baca-Garcia E (2012) Suicide attempters classification. Toward predictive models of suicidal behavior. Neurocomputing 92:3–8
15. Benton A, Mitchell M, Hovy D (2017) Multi-task learning for mental health using social media text. In EACL, ACL
16. Ji S, Long G, Pan S, Zhu T, Jiang J, Wang S (2019) Detecting suicidal ideation with data protection in online communities. DASFAA. Springer, Cham, pp 225–229
17. Pavalanathan U, De Choudhury M (2015) Identity management and mental health discourse in social media. In WWW, pages 315–321. ACM
18. Wang Y, Tang J, Li J, Li B, Wan Y, Mellina C, O'Hare N, Chang Y (2017) Understanding and discovering deliberate self-harm content in social media. In WWW, pages 93–102
19. Shuai HH, Shen CY, Yang DN, Lan YF, Lee WC, Yu PS, Chen MS (2016) Mining online social data for detecting social network mental disorders. In WWW, pages 275–285
20. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. In AAAI ICWSM
21. Li Y, Mihalcea R, Wilson SR (2018) Text-based detection and understanding of changes in mental health. In International Conference on Social Informatics, pages 176–188. Springer
22. Cao L, Zhang H, Feng L (2020) Building and using personal knowledge graph to improve suicidal ideation detection on social media. IEEE Trans Multimed. https://doi.org/10.1109/TMM.2020.3046867
23. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M (2014) Affective and content analysis of online depression communities. IEEE Trans Aff Comput 5(3):217–226
24. Ren F, Kang X, Quan C (2016) Examining accumulated emotional traits in suicide blogs with an emotion topic model. IEEE J Biomed Health Inform 20(5):1384–1396
25. Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J (2021) Deep learning based text classification: A comprehensive review. ACM Computing Surveys 54
26. Kim Y (2014) Convolutional neural networks for sentence classification. In EMNLP, pages 1746–1751
27. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
28. Lai S, Xu L, Liu K, Zhao J (2015) Recurrent convolutional neural networks for text classification. AAAI 333:2267–2273
29. Li Y, Pan Q, Wang S, Yang T, Cambria E (2018) A generative model for category text generation. Inf Sci 450:301–315
30. Zhao W, Peng H, Eger S, Cambria E, Yang M (2019) Towards scalable and reliable capsule networks for challenging NLP applications. In ACL, pages 1549–1559
31. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
32. Lin Z, Feng M, Santos CND, Yu M, Xiang B, Zhou B, Bengio Y (2017) A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130
33. Ma Y, Peng H, Cambria E (2018) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In AAAI, pages 5876–5883
34. Hamilton WL, Clark K, Leskovec J, Jurafsky D (2016) Inducing domain-specific sentiment lexicons from unlabeled corpora. In EMNLP, volume 2016, page 595
35. Santoro A, Raposo D, Barrett DG, Malinowski M, Pascanu R, Battaglia P, Lillicrap T (2017) A simple neural network module for relational reasoning. In NIPS, pages 4967–4976
36. Cambria E, Wang H, White B (2014) Guest editorial: big social data analysis. Know-Based Syst 69:1–2
37. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
38. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

39. Shing HC, Nair S, Zirikly A, Friedenberg M, Daumé III H, Resnik P (2018) Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on CLPsych*, pages 25–36
40. Zirikly A, Resnik P, Uzuner Ö, Hollingshead K (2019) CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In Proceedings of the Sixth Workshop on CLPsych
41. Coppersmith G, Dredze M, Harman C, Hollingshead K, Mitchell M (2015) Clpsych 2015 shared task: Depression and ptsd on twitter. In Proceedings of the 2nd Workshop on CLPsych, pages 31–39
42. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759
43. Pennington J, Socher R, Manning C (2014) GloVe: Global vectors for word representation. In *EMNLP*, pages 1532–1543

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.