# MAM: A Metaphor-based Approach for Mental Illness Detection

Dongyu Zhang[1], Nan Shi[1], Ciyuan Peng[2], Abdul Aziz[1], Wenhong Zhao[3], and Feng Xia[2][0000−0002−8324−1859]

[1] School of Software, Dalian University of Technology, Dalian 116620, China
[2] School of Engineering, IT and Physical Sciences, Federation University Australia, Ballarat, VIC 3353, Australia
f.xia@ieee.org
[3] Ultraprecison Machining Center, Zhejiang University of Technology, Hangzhou 310014, China

**Abstract.** Among the most disabling disorders, mental illness is one that affects millions of people across the world. Although a great deal of research has been done to prevent mental disorders, detecting mental illness in potential patients remains a considerable challenge. This paper proposes a novel metaphor-based approach (MAM) to determine whether a social media user has a mental disorder or not by classifying social media texts. We observe that the social media texts posted by people with mental illness often contain many implicit emotions that metaphors can express. Therefore, we extract these texts' metaphor features as the primary indicator for the text classification task. Our approach firstly proposes a CNN-RNN (Convolution Neural Network - Recurrent Neural Network) framework to enable the representations of long texts. The metaphor features are then applied to the attention mechanism for achieving the metaphorical emotions-based mental illness detection. Subsequently, compared with other works, our approach achieves creative results in the detection of mental illnesses. The recall scores of MAM on depression, anorexia, and suicide detection are the highest, with 0.50, 0.70, and 0.65, respectively. Furthermore, MAM has the best F1 scores on depression and anorexia detection tasks, with 0.51 and 0.71.

**Keywords:** mental illness · metaphor · attention model · text classification.

## 1 Introduction

According to the WHO survey, about 13% of people worldwide have a mental illness[4], and approximately 800,000 people suicide every year, with almost one death every 40 seconds[5]. Most people with mental illness are reluctant to share their feelings with others in real life and have a prevalent tendency to be

[4] `https://www.who.int/health-topics/mental-health`
[5] `http://www.who.int/data/gho/data/themes/mental-health`

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

alone. However, with the emergence of social media platforms, such as Facebook, WeChat and Twitter, etc., many mental disorder patients are willing to talk about their illnesses and share their feelings with strangers online. Thus, many reliable texts that reflect people's emotions and sentiments can be obtained from social media. Likewise, it becomes easy and possible to detect one's mental state by analyzing the text content posted by them.

In particular, many patients with mental illness tend to hide their feelings and emotions in the text through implicit means of expression, such as metaphor, rather than expressing them directly [7, 13, 15]. While sharing the feelings and sentiments, many people with mental illness often express their emotions through many different metaphorical languages, by which they intend to describe something other than what has been written on social media. For example, some people post the metaphorical expressions such as "My father is a monster" or "My life is a prison" frequently, these kind of expressions can help in detecting their hidden mental illnesses.

In general, there are many ongoing studies about mental illness detection via social media [17, 19]. Typically, text classification algorithms have broadly been applied for prediction of people's mental health state based on their social status in the form of text posted on social media platform. To improve text classification models' performance, additional information extracted from the texts needs to be as the learning features [23]. However, current research on mental health issues only focuses on extracting some explicit features such as topics to indicate the text classification task. The implicit emotional feelings expressed by metaphors and inherently related to mental illnesses are ignored. This paper proposes a metaphor-based approach to fill this gap, which takes advantage of metaphorical expressions in the texts to detect mental illnesses. We extract the metaphor features which reflect some implicit emotions of potential mental illness patients as the primary indicator of the text classification task. However, it should be noted that there are also some patients who do not use metaphors to express their emotions. Therefore, our approach only focuses on detecting the mental illness of the patients who write in a metaphorical way to express their feelings. Our approach is divided into three main processes: long text representation based on CNN-RNN (Convolution Neural Network - Recurrent Neural Network) framework, metaphor feature extraction, and attention mechanism based text classification.

Initially, we take the advantage of CNN-RNN framework to represent the textual data. Some studies on deep learning-based text classification have been done in this regard [3, 28]. Nevertheless, most of them only focus on short texts which contain hundreds of words. Social media texts contain thousands or millions of words, and it is hard for the ordinary models to achieve good performance on the social media text classification task. We apply the CNN-RNN framework to process long text data by extracting the information from words and sentences, respectively. Similar to TextCNN (the convolutional neural network for text) [6], We firstly apply CNN to extract word features from sentences and then obtain sentence representations. All the sentence representations in a text are processed

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

by the bidirectional Long Short Term Memory (Bi-LSTM) model to obtain text representation. Conspicuously, the CNN-RNN framework's application makes our approach more efficient on long text extraction tasks with lower memory usage.

Because of the inherent connection between metaphorical expression and mental health, we extract the metaphor features as the text classification task's primary indicator. To extract metaphor features, we apply the token-level metaphor identification method RNN_MHCA (Recurrent Neural Network Multi-Head Contextual Attention) [5, 16] to identify the metaphors from text. Significantly, for better performance of our approach, we extract sentence-metaphors and text-metaphor features respectively. Then, an attention mechanism is used to calculate attention weights based on the sentence-metaphor and text-metaphor features. Consequently, the attention mechanism integrates the text representation obtained from the CNN-RNN framework with the attention weights to get the text vectors for the text classification.

Finally, we test our model on the datasets eRisk [14] and CLPsych (The Computational Linguistics and Clinical Psychology) [1, 30]. The experimental results demonstrated that our model performs well.

The main contributions of this study are as follows:

– A novel metaphor-based approach is proposed to detect users' mental illness status with the help of social media platforms. Specifically, our approach classifies social media texts based on mental illness patients' metaphorical expressions expressed over social media platform.
– Basically the initial aim of our approach is to represent the long texts based on the CNN-RNN framework. Remarkably, our model gains the competitive results with processing long textual data from social media.

The remainder of this paper is organized as follows. Section 2 introduces related work on text classification and metaphor recognition. Section 3 presents details about the proposed metaphor-based attention model. Experiments and results are discussed in Section 4. Finally, Section 5 concludes the paper.

## 2 Related Work

Research on textual data, specifically text classification, has gained much attention in recent years. In particular, the most popular text classification methods focus on semantic feature extraction, and inner logic in context [27].

Bi-LSTM has been used by [28] to obtain the contextual representation of the words in the sentence. Then, in their model, CNN is employed on the encoder's output to obtain text features. However, CNN's output vector is treated as the calculation benchmark of the attention mechanism instead of the representation for classification. The contextual representation of each position is connected with the last hidden layer state in the recurrent network. Moreover, the text representation is the weighted sum of contextual representation accumulated through the attention mechanism. Likewise, Yao et al. [24] combined

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

the graph convolutional neural network with the text classification task. By converting documents and words into graph nodes, the graph neural network algorithm is successfully applied to text classification. They introduced the graph's adjacency matrix and degree matrix with stacked multiple graph convolutional network layers for multi-level extraction of information about the target node and surrounding neighbors.

Many text classification methods have been applied to mental illness detection tasks [17, 29]. In particular, a large number of them focus on social media text-based mental illness detection [19]. Hierarchical recurrent neural networks with attention mechanism are used to extract the specific expressions at word-level and sentence-level (e.g., I have depression) [9, 20]. Then, their model classifies the texts for mental illness detection based on these specific expressions. Considering the connection between words and categories of text [2, 12] learned and integrated the correlation information for text classification. Compared with methods based on deep learning, their method has a stronger interpretability and supports in incremental learning instead of retraining when adding new samples.

Yao et al. [23] directly trained patterns to capture useful phrases with word embedding technique. They combined these word embeddings with entity embeddings of the Unified Medical Language System and Concept Unique Identifiers as external knowledge. These embeddings are independently condensed through CNN and pooling layer, then connected for text classification after multi-layer perceptron. According to [10] texts' sentiment and topic are the leading indicators for mental illness detection. They obtained the emotional features by counting emotional words and topics obtained from the LDA (Latent Dirichlet Allocation) topic model. The contextual vectors, encoded by the RNN module, are concatenated with representations of sentiment or topic. Then the contextual vectors are inputted into relation networks to calculate relation vectors that are used in the attention mechanism to get text representation.

Metaphor identification is a linguistic metaphor processing task with concrete words [4, 25]. There are many methods for metaphor identification, such as Selectional Preference Violation (SPV) [4, 22] and Metaphor Identification Procedure(MIP) [21]. SPV depends on the semantic contrast between the target word and its context. MIP focuses on the contrast between the literal meaning and the contextual meaning of the target word. Nevertheless, metaphor identification has been treated as a sequence tagging task recently [16, 26]. Therefore, the deep learning methods, similar kind of approaches and the attention network have widely been applied for the tasks of metaphor identification [18].

Although metaphor identification research is extending in terms of development, few studies directly use metaphors as the text feature to achieve text classification, specifically, the text classification for mental illness detection. This paper considers the metaphor features in the texts as the primary indicators of text classification for mental illness detection.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

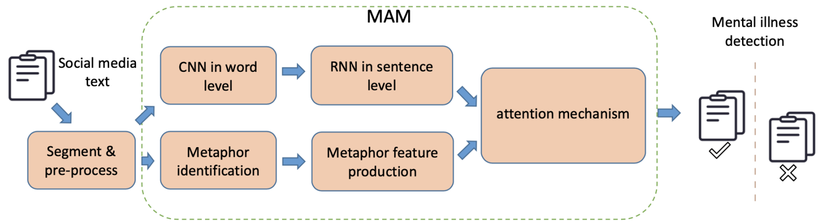


**Fig. 1.** Data processing in MAM.

## 3 Metaphor-based Attention Model

This section highlights the details of our approach MAM that achieves creative results on mental illness detection. There are three main processes of our approach MAM namely, the CNN-RNN framework based long text representation, metaphor feature extraction, and attention mechanism based text classification. The overall data processing of MAM approach is illustrated in Fig. 1. As it is clearly shown in Fig. 1, we first segment and pre-processed the social media textual data. Then, CNN is applied to process the words in the texts and passes the obtained sentence representation to RNN to process the sentence-level data. Therefore, the text representation can be obtained after the processing of RNN. On the other hand, we also extracted metaphor features by identifying metaphors and then producing the features from them. Lastly, an attention mechanism calculates attention weights based on the metaphor features and integrates the text representation with the attention weights to get the text vectors for text classification. In our approach, we classify the texts by calculating the texts' association with the labels (0: health control group, 1: positive group).

The entire architecture of MAM is shown in the Fig. 2. Each sentence is composed of words $\{x_1, x_2, ..., x_m\}$. Firstly, the word features in each sentence are extracted in the convolutional layer, and then we obtain sentence representations $\{c_1, c_2, ..., c_n\}$in the average pooling layer. Afterward, the sentence representations are encoded by bidirectional long short-term memory that captures the sequential contextual information in each sentence. Consequently, the text is represented as $\{h_1, h_2, ..., h_n\}$. Then, in the hidden layer, the attention weights $\{w_1, w_2, ..., w_n\}$ are calculated based on sentence-metaphor feature $\{m_1, m_2, ..., m_n\}$ and text-metaphor feature $t$. Finally, the text representation is integrated with the attention weights to obtain the text vectors for text classification.

### 3.1 CNN-RNN Framework for Long Text Representation

We first apply CNN to capture word features since CNN can notice the collocation between words in a sentence, i.e. abnormal word pair [11]. Taking a single $sample = (text, label)$ as an example of our model, the words $\{x_1, x_2, ..., x_m\}$ in

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

each sentence are embedded by 300d Glove vector. Each sentence's input is a matrix of $m \times 300$, where $m$ is the number of words in the sentence. The convolutional kernel $k$ performs the convolution operation with a window $x_{i:i+h-1}$ on the input matrix and produces the representation of each sentence as formulas:

$$c_{ki} = f(w_k \cdot x_{i:i+h-1} + b_k)$$

$$c_k = ave\_pool([c_{k1}, c_{k2}, ..., c_{k(m-h+1)}])$$

$$c_j = [c_1, c_2, ..., c_k].$$

Here, $x_{i:i+h-1}$ denotes a window of size $h \times 300$ consisting of row $i$ to row $i+h-1$ of the input matrix. $w_k$ is the weight matrix, and $b$ indicates the bias. $f$ is the activation function of the convolutional kernel $k$ which performs sequential scanning on the input matrix and stitches results to obtain $[c_{k1}, c_{k2}, ..., c_{k(m-h+1)}]$ as the feature maps. These feature maps are inputted to the average pooling layer that obtain the average value $c_k$ for the overall position collocation information. Then, the sentence representation $c_j$ is obtained by using various convolution kernels in different kernel size to convolve the sentence. Here, $j$ denotes the $j$-th sentence in the text.

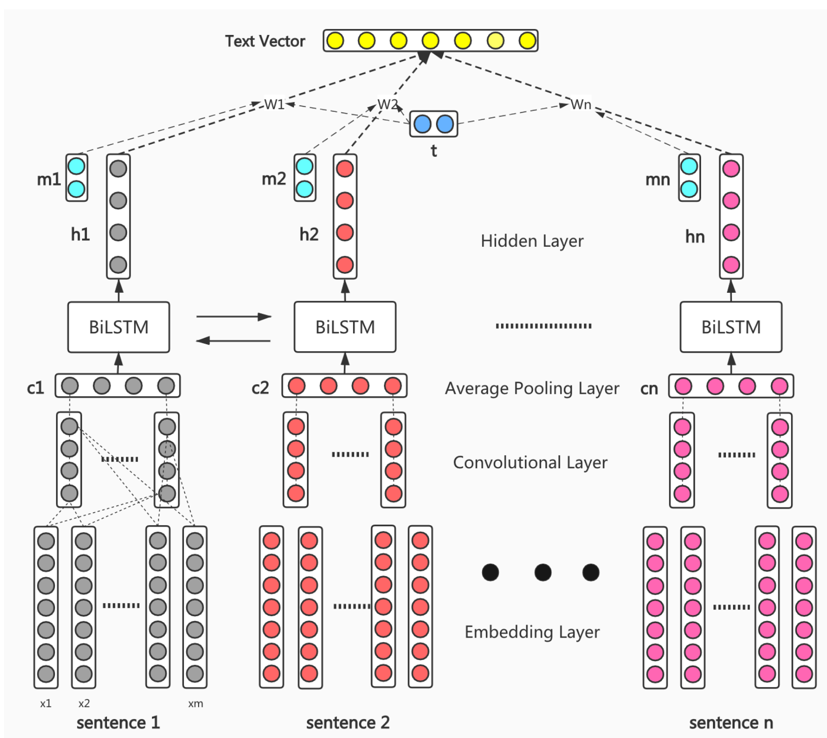


**Fig. 2.** The entire architecture of MAM.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

Because the normal RNN does not perform well when dealing with the long sequences. Moreover, the long short term memory (LSTM) improves the performance of RNN by introducing the gate mechanism and the cell state [8]. Therefore, we apply the Bi-LSTM to encode the sentence vectors based on the sequential and positional information between sentences in the text. The formulas of LSTM are shown as follows:

$$forget_j = \sigma(W_{forget} \cdot [h_{j-1}, c_j] + b_{forget})$$
$$input_j = \sigma(W_{input} \cdot [h_{j-1}, c_j] + b_{input})$$
$$\widetilde{C_j} = tanh(W_c \cdot [h_{j-1}, c_j] + b_c)$$
$$C_j = forget_j * C_{j-1} + input_j * \widetilde{C_j}$$
$$output_j = \sigma(W_{output} \cdot [h_{j-1}, c_j] + b_{output})$$
$$h_j = tanh(C_j) * output_j,$$

where, $W$ denotes the weight matrix, $b$ indicates the bias. $c_j$ is the $j$-th sentence representation obtained from the CNN model, and the hidden representation of the $j$-th sentence $h_j$ participates in the calculation as a part of the input. The cell state $C_j$ is used for recording the sequential state information in the $j$-th sentence, and the temporary cell state $\widetilde{C_j}$ is used to record the sequential state information temporarily. Gates $forgetj, inputj, outputj$ control what is forgotten, inputted and produced, respectively, and they are calculated by input vector $c_j$ and $h_{j-1}$, here, $h_{j-1}$ denotes the representation of the $(j-1)$th sentence. The new hidden representation $h_j$ and cell state $C_j$ will be employed in the next hidden layer.

Though the single direction structure of LSTM can not wholly extract all context in actual usage. Thus, we choose bidirectional LSTM, which can capture sequential information more effectively to process the data. In Bi-LSTM, the hidden sentence representation consists of the representations with two different directions. The formulas for calculating the sentence representation with different directions are as follows:

$$\overrightarrow{h_j} = hidden\_layer(c_j, h_{j-1})$$
$$\overleftarrow{h_j} = hidden\_layer(c_j, h_{j+1})$$
$$h_j = [\overrightarrow{h_j}, \overleftarrow{h_j}],$$

where, $\overrightarrow{h_j}$ indicates the hidden sentence representation with direction from left to right, and $\overleftarrow{h_j}$ is the hidden sentence representation with direction from right to left. Afterward, the sentence representation of $j$-th sentence is $h_j$ which consist of $\overrightarrow{h_j}$ and $\overleftarrow{h_j}$.

Finally, the text can be represented by the list of sentence representations $\{h_1, h_2, ..., h_n\}$. Here, $n$ is the number of sentences in the text.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

### 3.2 Metaphor Feature Extraction

People with mental illnesses use metaphors separately with various people, we extract metaphorical information as a selection standard of the contextual sentence representation. In terms of the extraction of metaphors, we use the token-level metaphor identification method RNN_MHCA to identify the metaphors. Although it cannot guarantee that all metaphors in the text can be exactly detected, because the identified metaphors directly affect the performance of our model, the accuracy of metaphor identification is high if our model performs well. RNN_MHCA compares a representation of the target word with its context that is captured by a bidirectional attention network. Then, according to the difference between the context and the target word, we can judge whether the target word has metaphorical meanings.Then, the target words with metaphorical meanings are considered as the metaphorical words in the text, called metaphors.

According to the identified metaphors, our model extracts sentence-metaphor and text-metaphor features, respectively, as the text classification task's primary indicator. We consider that people with mental illness use metaphors in different frequency, so the number of metaphors in a text is one of the important metaphor features. Furthermore, patients with different mental illnesses use metaphors in different forms. Therefore, we also consider that the POS (Parts-Of-Speech) tag of metaphors is one of the metaphor features. Because of the different ways, people with a mental health condition using metaphor are based on many aspects. We extract metaphor features based on the following factors:

- The number and proportion of metaphors
- The POS tag of metaphors, including noun, verb, and adjective, etc.
- The length and number of sentences

The number and proportion of obtained metaphors are considered the most important factors for extracting metaphor features from a text. Also, we consider the POS tags of metaphors obtained by NLTK[6] as one of the factors of metaphor features. In particular, for the extraction of text-metaphor features, the length and the number of total sentences in a text are also considered as the leading factors. Therefore, the sentence-metaphor features are mainly extracted based on the number and proportion of metaphors in one sentence and the metaphors' POS tags. For the text-metaphor features, they are not only extracted based on the number and the proportion of metaphors and the POS tags, but also based on the length and number of the sentences in the text.

### 3.3 Attention Mechanism for Text Classification

Compared with the methods adding metaphorical information as an additional feature directly, we add metaphor features to the representation and use the metaphor features as the text information selection benchmark. It makes the approach pay more attention to the sentence information, which has some characteristics in using the metaphors. In attention mechanism, the text-metaphor

[6] http://www.nltk.org

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

vector and the sentence-metaphor vector are transformed into 20 dimensional metaphor representations $t \in R^{1\times d}$ and $m_i \in R^{1\times d}$, respectively, through a fully connected layer. Here, $m_i$ means the metaphor vector of the $i$-th sentence in the text. $d$ is the dimension of the metaphor representation. Afterwards, the attention mechanism calculates the attention weights $w_i$ of the $i$-th sentence based on the metaphor representations. The calculation is as follows:

$$w_i = Softmax(t \cdot m_i^T).$$

Then, the hidden representation of the $i$-th sentence $h_i$ and the $i$-th sentence-metaphor vector $m_i$ are concatenated by a a fully-connected layer $L$. $w_i$ multiplies the sentence vector which is transformed from the concatenation of them. Consequently, the text vector $r$ can be obtained:

$$r = \sum_{i=1}^{n} w_i \cdot (L \cdot [h_i, m_i] + b_L),$$

here, $n$ is the number of sentences, $b_L$ denotes the bias. Lastly, the text vector $r$ is connected with output layer for classification:

$$p(la\hat{b}el|r,t) = \sigma(W_r \cdot [r,t] + b_r),$$

where, the *label* indicates the labels of the training data. Typically, the labels are denoted as 0 (the health control group) and 1 (the positive group). $W_r$ is the weight matrix, and $b_r$ denotes the bias.

## 4 Experiments

### 4.1 Datasets and Implementation

We have chosen two datasets for testing our model: eRisk and CLPsych. These two datasets are both constructed from social media texts and applied in shared tasks. We use the second edition of eRisk that contains two tasks - depression and anorexia and the suicide assessment dataset in the 2019 CLPsych shared task. For more details, please refer to [14, 30].

The dataset of eRisk serves as a collection of writings (posts or comments) from a set of social media users. The newest edition includes three types of characteristics for mental illnesses: depression, anorexia and self-harm. The eRisk dataset considers time sequence information by grouping social media text in chronological order. However, in our experiment, we do not consider the time factor and treat each chunk of a user as an individual sample with the same label to extend the contribution in experimental datasets. The CLPsych is a shared risk detection task for mental illnesses based on social media texts. It retains the texts' time information while the eRisk dataset groups the texts by time to meet the early detection tasks requirement. Based on the considerations of CLPsych data's reliability, we employ *Expert* dataset that was annotated by experts for

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

245 users who posted on SuicideWatch and 245 control users who did not. The annotation scale includes *a* - no risk, *b* - low risk, *c* - moderate risk and *d* - severe risk. According to the explanation of low risk: annotators do not think this person has a high risk of suicide, we divide *c* and *d* into positive data, and the rest is negative data. The *Expert* dataset is not divided into the training set and the test set. We use ten-fold cross-validation in the experiment.

For the validity of data, we remove the sentences with less than three words and the texts with less than two sentences from the datasets. The statistic of the datasets is shown in Table 1 (de: depression, an: anorexia).

**Table 1.** The statistics of the experimental dataset.

| | Train_de | Test_de | Train_an | Test_an | Suicide |
|---|---|---|---|---|---|
| The number of samples | 8810 | 8116 | 1513 | 3165 | 479 |
| Average tokens of sentence | 16.78 | 16.69 | 16.73 | 17.04 | 25.51 |
| Average sentence of sample | 132.96 | 169.17 | 128.08 | 118.62 | 298.69 |

During the experimental process, we set the batch size of 4, kernel size in the CNN layer of 2 and 3, and the number of filters is 200. The output dimension of Bi-LSTM and the fully-connected layer is 200. The RNN_MHCA is pre-trained on VUA dataset [21] for the process of metaphor feature extraction. Due to the data imbalance, the positive samples in all experiments have two times the negative samples' weight in the calculation of the loss.

### 4.2 Results

We compare the experimental results of MAM with some previous known research works, including Text-CNN, Bi-LSTM, the Bi-LSTM+Attention model, and the hierarchical RNN+Attention model. In Text-CNN, the feature maps obtained by the convolution kernel will be pooled and became the features used for classification [11]. Bi-LSTM is suitable for sequence problems, such as text and timing research, and can capture the context information. Bi-LSTM usually uses the output of the last position as a classification vector. The Bi-LSTM+Attention model adds an attention layer based on Bi-LSTM, and it calculates weights through the attention mechanism to integrate the information from every position of the text. The hierarchical RNN+Attention model firstly collects sentence representations from word features and then collects them into text features for classification. The experimental results are shown in Table 2, and all the metrics are based on the positive label.

It is clearly shown in Table 2 that Text-CNN achieves the best results in precision on depression and anorexia detection tasks, with 0.60 and 0.87. However, it has general experimental performance totally, typically in the recall of positive samples. The hierarchical RNN+Attention model achieves the highest precision of 0.93 and the highest F1 score of 0.68 in suicide detection tasks,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

**Table 2.** Experimental results on eRisk dataset.

| | | Depression | Anorexia | Suicide |
|---|---|---|---|---|
| Text-CNN | P | **0.60** | **0.87** | 0.84 |
| | R | 0.29 | 0.48 | 0.33 |
| | F1 | 0.39 | 0.61 | 0.48 |
| Bi-LSTM | P | 0.49 | 0.65 | 0.68 |
| | R | 0.38 | 0.61 | 0.44 |
| | F1 | 0.43 | 0.63 | 0.53 |
| Bi-LSTM+Att | P | 0.59 | 0.71 | 0.68 |
| | R | 0.40 | 0.66 | 0.58 |
| | F1 | 0.47 | 0.68 | 0.63 |
| H_RNN+Att | P | 0.59 | 0.74 | **0.93** |
| | R | 0.30 | 0.47 | 0.54 |
| | F1 | 0.40 | 0.57 | **0.68** |
| MAM | P | 0.52 | 0.72 | 0.70 |
| | R | **0.50** | **0.70** | **0.65** |
| | F1 | **0.51** | **0.71** | 0.67 |

but its performance is not so good on the other two datasets. In general, our approach MAM performs the best as compared to the other mentioned models. Despite the fact, our approach MAM generally performs very well when considering precision alone, its highest recall scores on all the mental illness detection tasks make the overall performance the most competitive. The recall scores of MAM on depression, anorexia, and suicide datasets are respectively 0.50, 0.70, and 0.65. Furthermore, MAM has the best F1 scores on depression and anorexia detection tasks, with 0.51 and 0.71. Moreover, MAM has less time and space consumption than the standard RNN-based methods in processing long text data and achieves competent results.

Compared with standard attention mechanisms, the attention mechanism in MAM does not let the model learn attention weights by itself but assigns a reference benchmark for the model. This mechanism does not capture the text's specific expressions but mines the implicit information in the text. It is based on the idea that patients with mental illnesses use metaphors differently from normal people. The results show that this attention mechanism is sufficient.

### 4.3 Ablation Study

In this part, we study the influence of several important parts of our approach on performance. Table 3 shows the F1 scores in ablation study.

**Metaphor identification method** We compare BERT with RNN_MHCA by replacing the original metaphor features with the metaphor features identified by BERT. To enlighten the study, as shown in Table 3 (Bert_metaphor), the F1

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

scores are 0.49 in depression detection, 0.68 in anorexia detection and 0.66 in suicide detection. These scores prove that RNN_MHCA can identify metaphors more effectively than BERT.

**RNN module** We remove the RNN module by directly combining the CNN output with metaphor features. The experimental results are shown in Table 3 (Remove RNN). The F1 scores are 0.45 in depression, 0.64 in anorexia, and 0.64 in suicide. Therefore, it proves that the RNN module is important for the better performance of our model.

**Metaphor features** We first remove the metaphor features in our model. As shown in Table 3 (Remove_metaphor), the experimental results (F1: 0.45, 0.64 and 0.55) show that removing the metaphor features will lead to a bad performance. Hence, we replace the metaphor features with the normal attention features in our model. As shown in Table 3 (Replace_metaphor), the experimental results (F1: 0.49, 0.69 and 0.66) are also not satisfactory. Thereby, it can achieve a creative performance to apply metaphor features into the text classification task for mental illness detection.

**Table 3.** The F1 scores in ablation study.

| | Depression | Anorexia | Suicide |
|---|---|---|---|
| MAM | **0.51** | **0.70** | **0.67** |
| Bert_metaphor | 0.49 | 0.68 | 0.66 |
| Remove RNN | 0.45 | 0.64 | 0.64 |
| Remove metaphor | 0.45 | 0.64 | 0.55 |
| Replace metaphor | 0.49 | 0.69 | 0.66 |

## 5 Conclusion

In this paper, we have proposed a novel metaphor-based approach (called MAM) to detect the mental illnesses of people who post their feelings on social media. We classify the social media texts based on their metaphors to detect the writers' mental health state. The CNN-RNN structure is proposed to achieve the representation of long texts. Metaphor features in the texts are extracted as the primary indicator of the text classification task. We apply an attention mechanism to calculate the attention weights based on the metaphorical features, and after that, it also integrates the text representation with the attention weights to get the text vectors for text classification. Experimental results show that MAM performs pretty well in depression, anorexia, and suicide detection.

In future work, we will deeply study the specific effect of metaphor features using bi-LSTM and various similar algorithms in mental illness detection, and demonstrate the specific inherent connection between metaphors and mental illness. We will compare other extraction frameworks to represent text features that are most suitable for metaphor information.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

## Acknowledgment

This work is partially supported by National Natural Science Foundation of China under Grants No. 62076051.

## References

1. Aguilera, J., Farías, D.I.H., Ortega-Mendoza, R.M., Montes-y Gómez, M.: Depression and anorexia detection in social media as a one-class classification problem. Applied Intelligence pp. 1–16 (2021)
2. Burdisso, S.G., Errecalde, M., Montes-y Gómez, M.: A text classification framework for simple and effective early depression detection over social media streams. Expert Systems with Applications **133**, 182–197 (2019)
3. Chen, T., Xu, R., He, Y., Wang, X.: Improving sentiment analysis via sentence type classification using bilstm-crf and cnn. Expert Systems with Applications **72**, 221–230 (2017)
4. Chen, X., Hai, Z., Wang, S., Li, D., Wang, C., Luan, H.: Metaphor identification: A contextual inconsistency based neural sequence labeling approach. Neurocomputing **428**, 268–279 (2021)
5. Gong, H., Gupta, K., Jain, A., Bhat, S.: Illinimet: Illinois system for metaphor detection with contextual and linguistic information. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 146–153 (2020)
6. Guo, B., Zhang, C., Liu, J., Ma, X.: Improving text classification with weighted word embeddings via a multi-channel textcnn model. Neurocomputing **363**, 366–374 (2019)
7. Gutierrez, E.D., Cecchi, G.A., Corcoran, C., Corlett, P.: Using automated metaphor identification to aid in detection and prediction of first-episode schizophrenia. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2923–2930 (2017)
8. Hua, Q., Qundong, S., Dingchao, J., Lei, G., Yanpeng, Z., Pengkang, L.: A character-level method for text classification. In: 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). pp. 402–406. IEEE (2018)
9. Ive, J., Gkotsis, G., Dutta, R., Stewart, R., Velupillai, S.: Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. pp. 69–77 (2018)
10. Ji, S., Li, X., Huang, Z., Cambria, E.: Suicidal ideation and mental disorder detection with attentive relation networks. arXiv preprint arXiv:2004.07601 (2020)
11. Kim, Y.: Convolutional Neural Networks for Sentence Classification. arXiv e-prints arXiv:1408.5882 (aug 2014)
12. Kumar, A., Sharma, K., Sharma, A.: Hierarchical deep neural network for mental stress state detection using iot based biomarkers. Pattern Recognition Letters (2021)
13. Llewellyn-Beardsley, J., Rennick-Egglestone, S., Callard, F., Crawford, P., Farkas, M., Hui, A., Manley, D., McGranahan, R., Pollock, K., Ramsay, A., et al.: Characteristics of mental health recovery narratives: systematic review and narrative synthesis. PloS one **14**(3), e0214678 (2019)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47

14. Losada, D.E., Crestani, F., Parapar, J.: Overview of erisk: early risk prediction on the internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 343–361. Springer (2018)
15. Magaña, D.: Cultural competence and metaphor in mental healthcare interactions: A linguistic perspective. Patient education and counseling **102**(12), 2192–2198 (2019)
16. Mao, R., Lin, C., Guerin, F.: End-to-end sequential metaphor identification inspired by linguistic theories. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3888–3898 (2019)
17. Preotiuc-Pietro, D., Sap, M., Schwartz, H.A., Ungar, L.H.: Mental illness detection at the world well-being project for the clpsych 2015 shared task. In: CLPsych@ HLT-NAACL. pp. 40–45 (2015)
18. Rivera, A.T., Oliver, A., Climent, S., Coll-Florit, M.: Neural metaphor detection with a residual bilstm-crf model. In: Proceedings of the Second Workshop on Figurative Language Processing. pp. 197–203 (2020)
19. Saravia, E., Chang, C.H., De Lorenzo, R.J., Chen, Y.S.: Midas: Mental illness detection and analysis via social media. In: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). pp. 1418–1421. IEEE (2016)
20. Sekulić, I., Strube, M.: Adapting deep learning methods for mental health prediction on social media. arXiv preprint arXiv:2003.07634 (2020)
21. Steen, G.: A method for linguistic metaphor identification: From MIP to MIPVU, vol. 14. John Benjamins Publishing (2010)
22. Wilks, Y.: A preferential, pattern-seeking, semantics for natural language inference. In: Words and Intelligence I, pp. 83–102. Springer (2007)
23. Yao, L., Mao, C., Luo, Y.: Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. BMC medical informatics and decision making **19**(3), 71 (2019)
24. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 7370–7377 (2019), https://doi.org/10.1609/aaai.v33i01.33017370
25. Zhang, D., Lin, H., Liu, X., Zhang, H., Zhang, S.: Combining the attention network and semantic representation for chinese verb metaphor identification. IEEE Access **7**, 137103–137110 (2019)
26. Zhang, D., Zhang, M., Peng, C., Jung, J.J., Xia, F.: Metaphor research in the 21st century: A bibliographic analysis. Computer Science and Information Systems **18**(1), 303–321 (2021)
27. Zhang, P., Huang, X., Wang, Y., Jiang, C., He, S., Wang, H.: Semantic similarity computing model based on multi model fine-grained nonlinear fusion. IEEE Access **9**, 8433–8443 (2021)
28. Zheng, J., Zheng, L.: A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification. IEEE Access **7**, 106673–106685 (2019)
29. Zhong, B., Huang, Y., Liu, Q.: Mental health toll from the coronavirus: Social media usage reveals wuhan residents' depression and secondary trauma in the covid-19 outbreak. Computers in human behavior **114**, 106524 (2021)
30. Zirikly, A., Resnik, P., Uzuner, Ö., Hollingshead, K.: CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. pp. 24–33. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/W19-3003, https://www.aclweb.org/anthology/W19-3003

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_47