Pattern Recognition Letters 157 (2022) 135–143

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Hyper-graph-based attention curriculum learning using a lexical algorithm for mental health

Usman Ahmed[a], Jerry Chun-Wei Lin[a,*], Gautam Srivastava[b,c]

[a] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
[b] Department of Mathematics and Computer Science, Brandon University, Brandon, Canada
[c] Research Centre for Interneural Computing, China Medical University, Taichung, Taiwan

ARTICLE INFO

Article history:
Received 17 December 2021
Revised 2 March 2022
Accepted 19 March 2022
Available online 26 March 2022

Edited by: Maria De Marsico

MSC:
68T50
68U15
68U20
68T30

Keywords:
Internet-delivered interventions
Word sense identification
Text clustering
Adaptive treatments
NLP

ABSTRACT

In this paper, we propose a structure hypergraph and an emotional lexicon for word representation. Our method can solve problems related to vocabulary size, grammatical representation of words, and the lack of an emotional lexicon. Natural Language Processing (NLP) and attention-based curriculum learning are then used in the developed model. The goal is to achieve semantic word representations using a graph model. Later, embedding is used to label the text using clinical procedures. The experimental results show the emotional word representation with the structure hypergraph. The bidirectional Long Short Term Memory (LSTM) architecture with an attention mechanism achieved a Receiver Operating Characteristic (ROC) value of 0.96. The learning method can help psychiatrists in note taking and contributes to the detection rate of depression symptoms.

1. Introduction

According to WHO [1], depression is a serious problem that is one of the most disabling diseases in the world. Over 264 million people are affected by depressive disorders worldwide. In most cases, depression goes untreated because of a lack of verbal interaction and trust [2]. More than 800,000 people die each year from depression due to suicide [1]. Among other causes, suicide is the leading cause of death among people ages 15–29 and 76% and 85% in middle-income countries because they do not receive treatment for their condition. In addition to personal problems, lack of resources, untrained healthcare providers, social stigma, and timely intervention create further barriers to early detection [1]. People are often embarrassed about not having a stable mental state due to anxiety and shyness. This problem persists when the patient undergoes a probing examination of their psychological pain [3]. As a result, patients with depression may not undergo further treatment to make their current physical problems manageable.

Internet-delivered Psychological Treatment (IDPT) helps to cope with psychological problems by using fewer resources [4]. It mainly uses a tunnel-based solution that is inflexible and not interoperable [5]. The model lacks adaptability due to adherence and more dropouts [6]. Interventions need to consider user adoption over time. Adoption can be achieved with an IDPT system that takes into account the user's behavioural assessment and emotional interactions. However, personalized user behavior involves different preferences depending on the society, environment, and cultural symptoms of mental health [5].

1.1. Motivation

The COVID-19 pandemic has also exacerbated mental health issues worldwide. According to a report from WHO[1], 93% of countries worldwide have experienced an increase in mental health problems. Physiological stressors for people have increased due to

* Corresponding author.
E-mail addresses: Usman.Ahmed@hvl.no (U. Ahmed), chun-wei.lin@hvl.no, jerrylin@ieee.org (J.C.-W. Lin), SRIVASTAVAG@brandonu.ca (G. Srivastava).

[1] shorturl.at/enwTV.

https://doi.org/10.1016/j.patrec.2022.03.018

lockdown, which includes fear of disease and uncertainty about the future of humanity [7]. Social isolation, lack of interactive activities, educational uncertainty, and irregular work schedules have led to much higher rates of emotional stress. Fear of illness, lack of protective equipment, social isolation, and working in a highly stressful atmosphere have contributed to anxiety and depression among healthcare workers. Overall, the number of depression cases during the pandemic is significantly higher than anyone could have expected [8]. As the world moves toward online systems in many different areas, Internet forums and social media platforms allow people to connect [9]. Members of our global society have become accustomed to interacting online due to the pandemic [10]. Online detection of depression can help identify individuals at high risk for mental health. Timely medication treatment can then help improve overall well-being [11].

1.2. Contribution

In this paper, we propose a depression symptom extraction technique using Natural Language Processing (NLP) and attention-based curriculum learning. We proposed the semantic vectors based on an emotion-driven context extraction technique and a structural hypergraph. The method separates the important boundary elements from the unlabeled text and then incorporates them into the curriculum learning mechanism for the curriculum. The method updates the model training with new instances. The cycles are continuous until the optimal solution is reached and the pool of unlabeled text is included in the training set. The research objective is to expand the text information by increasing the knowledge of the learning method over time. The proposed method can help reduce the data annotation tasks and generalize the learning system. The semantic vectors and synonym expansion in the graph network help to achieve high accuracy without reducing the results of data annotation.

2. Related work

Chen et al. [12] examined data from Twitter to identify mental health problems. According to McDonnell 2020 and Chen 2020, the Linguistic Inquiry Word Count (LIWC) compares experimental and control groups. Two linguistic models are used to analyze word probability, e.g., (1) a unigram model and (2) a character-based 5-gram model to evaluate 5-character sequences. Nguyen et al. [13] explore a novel feature space representation in a neural network. The learned features represent the conditional probability distribution of the input vectors. Numerous architectures are proposed for application-specific domains.

Research addresses how to make empathic improvements to user experiences. Then, user expressions can be used to activate game and entertainment events in a cinematic approach to generate dynamic application behaviour. The method uses the realism factor to classify emotions. Multimodal-Adaptive Hierarchical Network (MAHN) uses the hierarchical recurrent neural network and an information modulation module between the hierarchical structure [14]. To increase the generalization ability of our model, we use a multi-task training technique to jointly maximize the BPR loss and the reconstruction loss of multimodal data. We conducted experiments on two real public datasets and the results show that our model outperforms the others. Another method uses a particle swarm optimization algorithm with linear weighted approaches [15]. The method helps to improve reliability and diversity. The model improves the performance of the time series of COVID-19 epidemic data and helps in the prevention and control of the COVID-19 epidemic.

The Gray Wolf Optimization (GWO) method was used [16]. The use of GWO allows the selection of optimal parameters for training the DNN model. Data standardization of diabetic retinopathy instances is performed using a standard scaling normalization approach, dimensionality reduction with PCA, selection of appropriate hyperparameters with GWO, and training of the data set with a DNN model. Model output was performed using Support Vector Machine (SVM), Naïve Bayes Classifier, decision tree, and XGBoost. The proposed approach uses noun-based filtering and a word ranking method that improves the performance of the text classification system. The frequency and distribution of the data are used to eliminate unnecessary attributes in the first step. Then, noun ranking is performed to improve the weighting of words. The irrelevant words are deleted in the second step based on their association with the output class. The method is compared with other feature extraction methods, including frequency-inverse document frequency, balanced accuracy measure, GINI index, information gain, and chi-square. The experimental results demonstrate the strength of the proposed algorithm.

Modern networks add hidden layers to the embedding that increase the predictive capacity of the model. As can be seen in [17], neural networks are classified according to the number of hidden layers, the type of layers, their shape, and the connections between them. Wainberg et al. [18] introduced the z technique for extracting higher-dimensional features from tabular data. Their convolutional neural network (CNN) is trained on the image pixels to learn feature embeddings. The network benefits from the translation-invariant pixels. In [19], a recurrent neural network (RNN) architecture was developed and applied to NLP applications with sequential data. The encoder-decoder architecture uses the RNN model to encode and decode the fixed-length sequence vector.

The model described above extracts text features or embedding methods based on statistical and deep learning. However, semi-supervised learning requires labels and training the model with adaptive output classes. In this study, we used the natural language processing method to build a tool for extracting depression symptoms from texts written by patients. We used the graph embedding method to extract, identify and visualize the graph attention method. Mainly, patients express mental health problems in their interaction with the system. Recognizing these sequences of communication patterns can help psychiatrists extract factors that cause mental and emotional unrest. The online interactive tool (ICT) provided can help provide contextual information and visualizations for appropriate mental health prevention interventions.

3. Hyper-graph based attention curriculum learning (HACL)

This paper proposes a graph attention embedding approach to efficiently and effectively identify depressive symptoms. As shown in Fig. 1, we use cosine similarity to generate symptom ratings after embedding. The latent space is constructed using a structure-aware graph model. Extended lexicons are constructed using extensive knowledge and graph embedding. This research aims to assist and improve psychiatrists' note taking based on graph attention networks. Each symptom group is labeled based on the frequency of patient texts.

3.1. Psychometric questionnaires (PQ)

The proposed technique uses the standard PHQ-9 questionnaire to collect texts written by patients [20]. PHQ-9 is a widely used method for assessing depressive symptoms. The PHQ-9 approach helps identify nine unique behavioural patterns included in the Diagnostic and Statistical Manual of Mental Disorders 5 (DSM-V)[2].

[2] https://www.psychiatry.org/psychiatrists/practice/dsm.

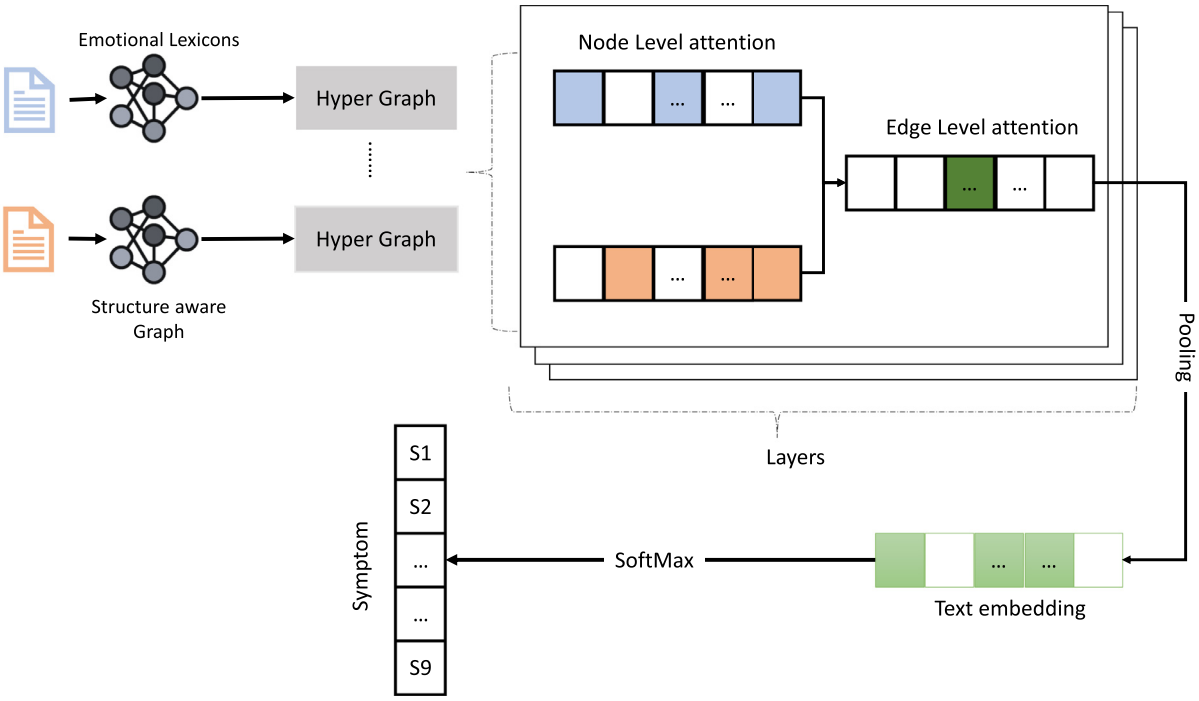


Fig. 1. A workflow architecture using the emotional lexicon and structure aware hyper graph.

The nine symptoms in the PHQ-9 are then grouped into different disorders, for example, sleeping, interest, concentration, and eating disorders, and example document[3]. Many studies have used an approach to assign depression scores. The solution was then tested on the PHQ-9 to measure the patient's depression score [21]. The other method includes *PHQ-2*, which contains 2 items, and *PHQ-15*, which contains 15 somatic symptoms [22] It is a complicated task to assess and diagnose mental health according to ICD10 [23] guidelines. The patient's mental health depends on several factors, such as family culture, previous treatments, society, childhood memories, work-life, and daily routines. Therefore, during screening and history taking, psychiatrists closely examine the factors that trigger the mental condition.

Generally, psychiatrists make lists of triggering events in patients and highlight the most important points. Based on the points recorded, psychiatrists recommend some exercises. At a second visit, psychiatrists use the standard procedure of a questionnaire based on the PHQ-9. The test is used to assess the clinical state of the patient's mental health. The test consists of different questionnaire schemes that include the type of symptoms, cause and treatment. A score is assigned by summing the frequency of symptoms. The resulting score describes the intensity of the mental health issues. For example, there are nine different symptoms; each symptom is further classified into mild, moderate, or severe conditions. The approach is called the "Clinical Symptom Elicitation Process" (CSEP) [23]. The psychiatrist generates a rating score following the questionnaire assessment. The rating score reflects the patient's level of depression. In the typical CSEP protocol, the psychiatrist asks the questions for each category and reviews the patient's responses to assign a frequency to the category, e.g., score0: never, b) score1: many days, c) score2: more than half of the days, and d) score3: virtually every day.

As shown in Fig. 2, the embedding discussed in Section 3.2 is used for the data labeling task. Starting from two texts, *PQ -9 questionnaire* and *patient authorized text*, the proposed model transforms them into two vectors, **t** and **e**, where both vectors are sentiment latent representations. For semantic expansion of *PQ-9 questionnaire*, we used WordNet-based seed term generation [21]. Then, the similarity with the vector *patient authorized* **e** with all symptoms of the extended latent representation (S1-S9) was measured. The output of the model is the multiclass prediction into nine different symptoms based on the latent similarity.

3.2. Structural-aware hypergraph

Using the example of the text written by a patient, we have interpreted it as a sequential directed acyclic graph G in which word points represent place and time. In this study, the edge between each node represents a sequence of words that occur sequentially, with the structure represented by the $\mathcal{D} = \{d_1, d_2, d_3, d_4\}$ word nodes. We used the hypergraph to represent the nodes in the graph. The study groups words based on their semantic meaning as a hypergraph, as shown in Fig. 3. The hyperedges have their attention network called nodes and the edges have their attention network called edge level attention. The hyperedges are grouped for sentence embedding using the averaging method and are called hyperedge vertices.

In hypergraph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, \ldots, v_n\}$ represents the set of nodes connected with two or more nodes (i.e., $\sigma(e) \geq 2$). The structure of the hypergraph G can be represented by an incidence matrix $\mathbf{I} \in \mathbb{D}^{n \times m}$, whose entries are defined as follows:

$$\mathbf{I}_{xy} = \begin{cases} 1, & \text{if } v_x \in e_y \\ 0, & \text{if } v_x \notin e_y \end{cases} \tag{1}$$

Each node in G is a d-dimensional attribute vector. Therefore, node in the graph is defined as vector $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^{\mathrm{T}} \in \mathbb{D}^{n \times d}$. We can represent the graph $G = (\mathbf{I}, \mathbf{X})$ which represents the hypergraph. We used the one hot vector by using the structural word model to preserve the sequential nature of the word in the patience authored texts.

[3] https://www.uspreventiveservicestaskforce.org/Home/GetFileByID/218.

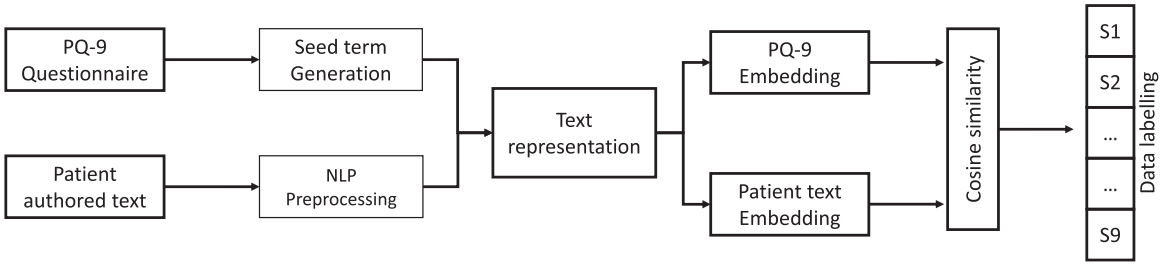


Fig. 2. A workflow data labeling using the text representations.

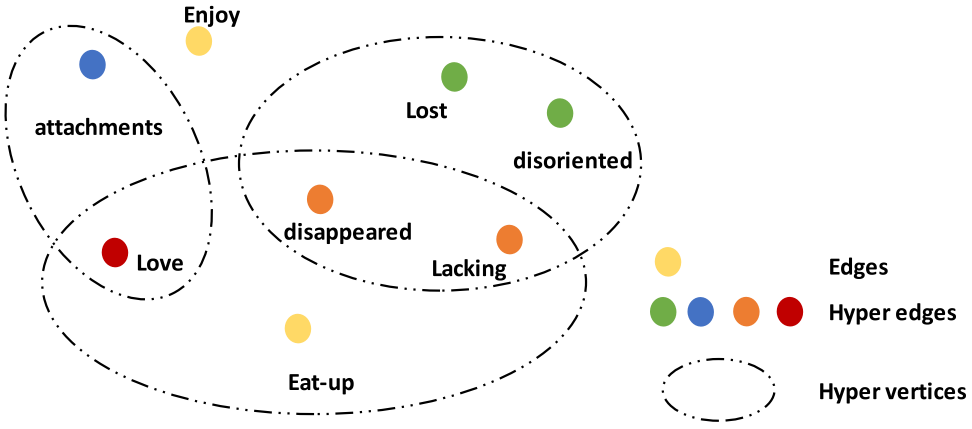


Fig. 3. A hyper-graph construction where edge and vertices are grouped by semantic representations.

3.3. Structural graph embedding

After the word-level modeling, the nodes are represented by latent learning representations. We used the ordered degree sequence for a collection of nodes $S \subset V$. By $T_k(x)$ nodes, we denote the number of hops between k points in G. For example, $T_1(x)$ represents the set of neighbors of vertex x when its distance is set as 1. $T_k(x)$ denotes the extension of the structure of nodes at a distance of k. When the ordered degree sequences of the two-word nodes x and y are compared (two nodes in the vertex network). The structural distance between x and y is denoted by $f_k(x, y)$. We consider their neighborhoods as k (all words that are within a distance of k. Eq. (2) defines the function as follows.

$$f_k(x,y) = f_{k-1}(x,y) + g(s(T_k(x)), s(T_k(y))) \quad k \geq 0 \text{ and } |T_k(x)|, |T_k(y)| > 0 \tag{2}$$

Eq. (2) is defined only when both x and y have an edge at a distance k. The $f_k(x, y)$ function may be used to determine the distance between ordered degree sequences, and k helps to compute the degree sequences of nodes that are at the same distance from x and y using structure growth at a distance. We calculated the distance between two ordered degree sequences using Dynamic Time Warping (DTW). This approach enables the extraction of usable distances that are more tolerant of sequences of varying lengths and the loose compression of sequence patterns [21,24]. DTW helps determine the optimal alignment of the growth sequences of the words x and y. Given a distance function $d(x, y)$ for each element in the sequences, the DTW aligns the sequences such that the sum of distances between matching elements is minimized [24]. We chose to use the distance function given in Eq. (3) because the growth of the trajectory is represented by the degree sequences of a node with its neighbors.

$$d(a,b) = \frac{\max(x,y)}{\min(x,y)} - 1 \tag{3}$$

There is a zero distance between two identical nodes with ordered sequences ($x = y$ and $d(x, y) = 0$). The multilayer weighted network encodes the nodes as word sequences to construct contexts. The structure node $G = (V, L)$ is associated with the diameter of $k*$ hops. We define the multilayer graph by examining the k hop neighbors of the node. The function defined above is used to assign the weights to the nodes. Eq. (4) defines the edge weight of a layer as $(n2)$. The technique is to label data in the manner described in Fig. 2.

$$w_k(x,y) = e^{-f_k(x,y)} \tag{4}$$

For weighted edges, we have n_k^* vertices and at most $k^*(n2) + 2n(k^* - 1)$. A multilayer graph generates contextual information about the order of words. No labeling information was required to determine structural similarity based on the nodes. We used a biased random walk to traverse the multilayer network, making random selections and weighted sequences. The random walk first decides with probability $(q > 0)$ whether to move through the layers or stay in the current layer. Eq. (5) gives the probability that a node u to a node v in the layer k stays in the current layer.

$$p_k(x,y) = \frac{e^{-f_k(x,y)}}{Z_k(x)} \tag{5}$$

in which $Z_k(x)$ represents the normalization factor for vertex x in layer k. It can thus be defined by Eq. (6).

$$Z_k(x) = \sum_{v \in V} e^{-f_k(x,y)} \tag{6}$$

3.4. Binary pairwise classification model

A deep neural network was trained with binary pairings. Gated units are used in a recurrent neural network. In sequential tasks, the LSTM network for attention positioning preserved long-term memory. For each element, we calculated the average value. Eq. (7) is used to derive the learning function F for the

relation R.

$$F(\boldsymbol{v}^x, \boldsymbol{v}^y) = \begin{cases} 1 & (\text{if } \boldsymbol{v}^x \text{ and } \boldsymbol{v}^y \text{ satisfy } R) \\ -1 & (\text{otherwise}) \end{cases} \tag{7}$$

The variable R indicates the degree of similarity between sentences in PHQ-9 questions. If the word order matches the normal words, they are classified as normal text; otherwise, they are classified as emotional. Each node in this study consists of unique word sequences. We obtain the embedding for the discussion in the previous section. Then we represented a group of nodes using the average embedding approach. The binary pairwise classification model averages the label features. Then, using cosine similarity, we group the structurally similar examples. We used a degree-based optimization approach to minimize the search space and run a larger network [24]. $g(\mathbf{v}_x, \mathbf{v}_y; \mathbf{w})$ can be expressed in terms of Eqs. 8 and 9, respectively.

$$g(\mathbf{v}_x, \mathbf{v}_y; \mathbf{w}) = f(\mathbf{v}_x; \mathbf{w}) \cdot f(\mathbf{v}_y; \mathbf{w}) \tag{8}$$

$$R_{xy} := \begin{cases} 1, & \text{if } \mathbf{l}_x \cdot \mathbf{l}_y \geq u(\lambda) \\ 0, & \text{if } \mathbf{l}_i \cdot \mathbf{l}_y < l(\lambda), \quad i, j = [1, \cdots, n] \\ \text{None}, & \text{otherwise} \end{cases} \tag{9}$$

where $u(\lambda)$ and $l(\lambda)$ are designed to select the similar and dissimilar samples respectively. "None" denotes the absence of exercise samples $(\mathbf{v}_x, \mathbf{v}_y, R_{xy})$. By increasing the sample size in each batch, we want to limit the formation of clusters by curriculum learning [25]. The reason is that extremely similar phrases have a very high probability of being selected in the training samples. Then, the RNN proceeds to identify the labels for the optimal features by gradually taking a batch of complex data, with *lambda* steadily increasing during the clustering process. Moreover, $u(\lambda) = l(\lambda)$ holds only when all samples are used for training. Each step is explained in detail by the Algorithm 1, where a phrase with node information is provided, the training embeddings f_w and λ are generated using the gradient values (Algorithm 1, input). Also, $u(\lambda)$ and $l(\lambda)$ are sample selection techniques (Algorithm 1, input). Then, we extract lexicons using m samples (Algorithm 1, lines 2–6). The emotion lexicon is trained using the word2vec model with window size = 2. The emotion lexicon helps to convert the nodes into a vector representation and combine them later for the hyper-graph and attention network (Algorithm 1, line 8). We insert the word structure (Algorithm 1, line 9) into the text. Then we implement the pooling algorithm for the attention network as mentioned in Section 3.4 (Algorithm 1, line 10). Then, a small batch with average embedding is selected (Algorithm 1, line 11–13), and the similarity is calculated to obtain the labels of the sentences (Algorithm 1, line 12). Then, we update the gradient technique (Algorithm 1, line 13). We used the `argmax` technique to obtain the cluster classes for the output points of the test patterns (Algorithm 1, lines 12–13).

Algorithm 1 Curriculum learning with graph attention embedding.

INPUT: $T = \{t_i\}_{i=1}^{n}$, λ, $u(\lambda)$, $l(\lambda)$, m. n represents number of words, w represents embedding size, m represents numbers of instances per batch.
OUTPUT: Symptoms label c_i of $t_i \in Questionare$.

1: **while** $K \leq \{1, 2, \dot{,} \frac{n}{m}\}$ **do**
2: Select training samples from T;
3: **foreach** $synonym = 1, T\epsilon\ w$ **do**
4: $terms \leftarrow wordnet.hyperonym(w)$;
5: $terms \leftarrow wordnet.hyponym(w)$;
6: $terms \leftarrow wordnet.antonyms$;
7: **end foreach**
8: $Emotionallexicon \leftarrow word2vec_{dep}(vocabulary, corpus, window = 2)$;
9: $Struture_{graph} \leftarrow Structure_{hypergraph}()$;
10: $Attention_{PositionalEncoder} \leftarrow (Emotionallexicon,\ Struture_{graph})$;
11: Apply pooling strategy by embeddings;
12: Determine similarity by utilizing Eqs.~8 and~9;
13: Update λ by the GD algorithm; ▷ GD is gradient descent
14: **end while**
15: **while** $T_i \in Instance$ **do**
16: $\{l_i\} = F(T_i; w)$;
17: $\{c_i\} = argmax_h(l_{ih})$;
18: **end while**
19: **Return** Symtons label c_i.

4. Experimental results and analysis

This study used two feature extraction methods based on (1) the emotion-based lexicon and (2) the structure-aware graph model. Both models used a glove vector of dimension 300 for vectorization. The emotion-based lexicon uses embedding to convert the text into node vectors in nine symptom lexicons. The structure embedding model then uses the hypergraph to extract word-based node patterns. The trained embedding is then used to label the text based on questions.

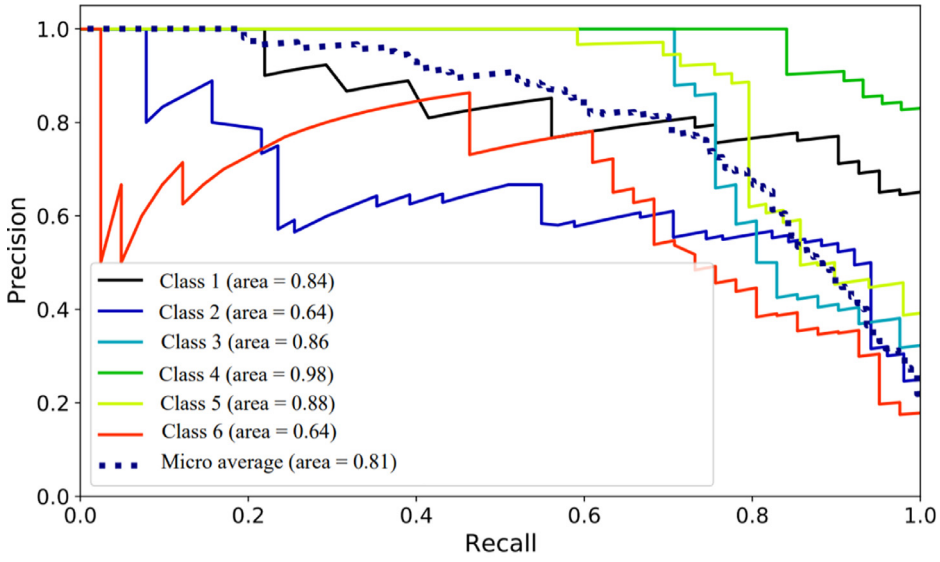
To categorize the nine particular symptoms in this study, we used a dataset from a variety of Internet forums and websites [5]. An entropy-based technique extends the learned knowledge and ensures that unusual occurrences do not affect the proposed system. The labelling is determined on a nine-point PHQ-9 scale, where 0 represents no depression, 1 represents mild depression, 2 represents moderate depression, and 3 represents severe depression [5]. We convert the label to a binary class for each symptom, where 0 represents no symptoms and 1 represents the presence of symptoms.

4.1. Model designation

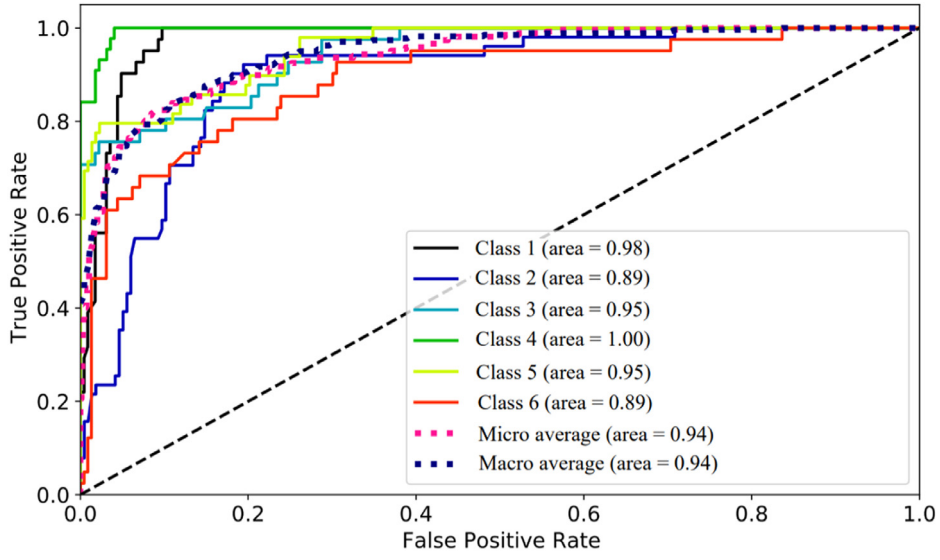
We used feed-forward neural networks as a baseline in the experiments. The moderate approach was used to keep the comment lengths consistent in size. The model consists of $(30, 20, 10)$ hidden layers with a ReLU activation function [21]. The goal of the method is to achieve high performance in classifying nine different symptoms using the hypergraph and the emotional lexicon. The final layer consists of a nine-link sigmoid function. The loss function is the cross-entropy function.

We used a recurrent neural network with gated units to achieve better performance on sequential tasks. In long-term memory (LSTM), the architecture takes the embedding sequence from the beginning and the end. Then the method sets one parameter for forward roll and one for backward roll. Therefore, the position encoders have two states, i.e., input and output. In addition to the LSTM units, we used the attention position layer [21]. For regularization and overfitting reasons, we set the dropout ratio for the hidden LSTM layers to 0.5. The sequence of a hypergraph and the emotional lexicon are kept separate to take advantage of the importance of words [26]. The attention network helps to extract structural information to select the important word in the attention layer.

The baseline model reaches the ROC of 0.94. This shows that the proposed feature extractor can contribute to high accuracy, as shown in Fig. 4. However, the model cannot capture unique patterns between classes when you look at the curve for accuracy and recall per class. The model achieves the 0.81 precision-recall curves. The model curve fluctuates at different thresholds. Therefore, further improvement is needed to achieve better performance. The depression data is a sequential prediction task where word sequences have relative importance. Therefore, an architec-



(a) Precision Recall Curve



(b) ROC curve

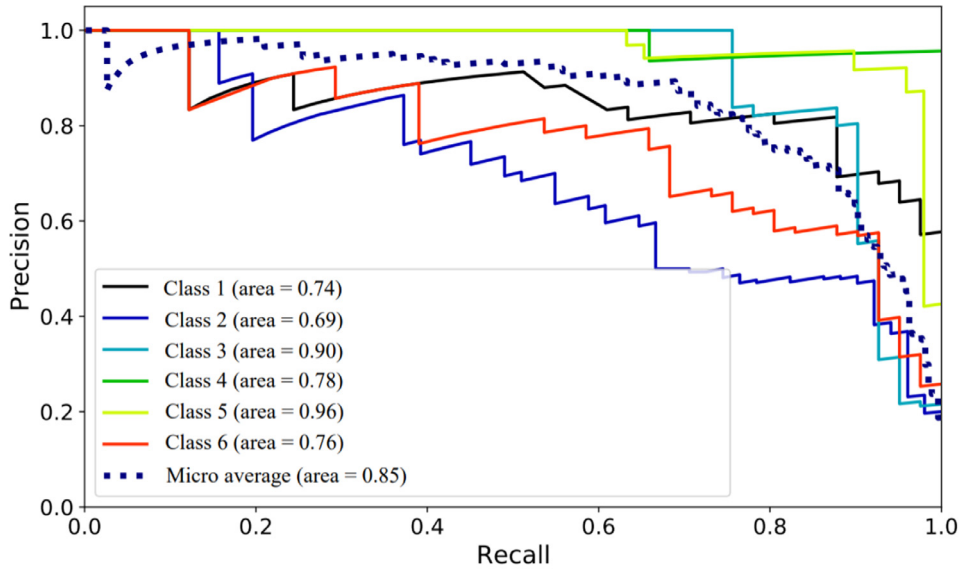
Fig. 4. Baseline model classification result.

ture that favors sequence and information storage can achieve high performance.

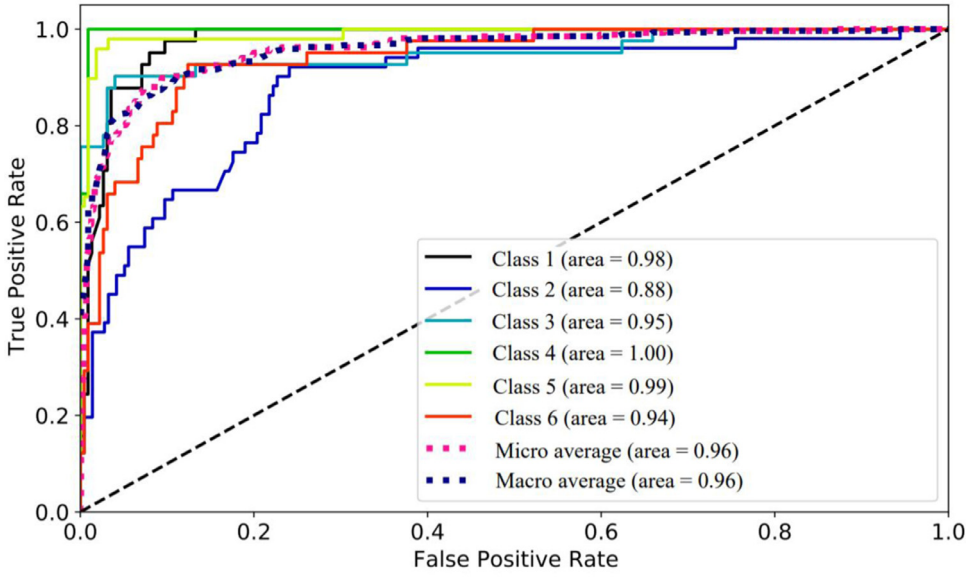
The LSTM network achieves relatively high performance compared to the baseline model but also exhibits a similar fluctuation problem as shown in Fig. 5. This indicates that the cross-class learning of the model is not optimal, as it improves from 0.81 to 0.85. The complex nature of LSTM cells has issues with the gradient disappearing as it moves from one cell to another. In architecture, hyper-tuning and longer run times can help reduce this problem. The network should run longer to achieve performance close to human levels. LSTM prefers initialization with small weights and follows the pattern of the baseline model.

As can be seen in Fig. 6, the bidirectional model achieved a high level of performance. The model ran in two directions based on input and output. The model's ability to learn sequence tasks is supported by the two separate RNN models. The testing set has the fewest errors, while the curve represents the upper corner. This shows the low rate of false positives and false negatives. The forward and backward sweep of the BILSTM model helps to maintain long dependencies that support long-term memory.

Positional attention with bidirectional LSTM is shown in Fig. 7. The model was used to determine the position vector to understand the correlated high-quality words. The model produces the fewest errors. The Receiver Operating Characteristic (ROC) value is 0.96 and the Precision-Recall curve is 0.86, respectively. This shows the high rate of true positives. The model includes the most important terms that contribute most to the categorization. By focusing on important words, the network helps reduce the cost of vector computation. The attention network can be trained with a directional coder to find significant words. The complexity of individual mental health data. In addition, the expanded vocabulary and grammatical combinations could help to strengthen the attention network.



(a) Precision Recall Curve



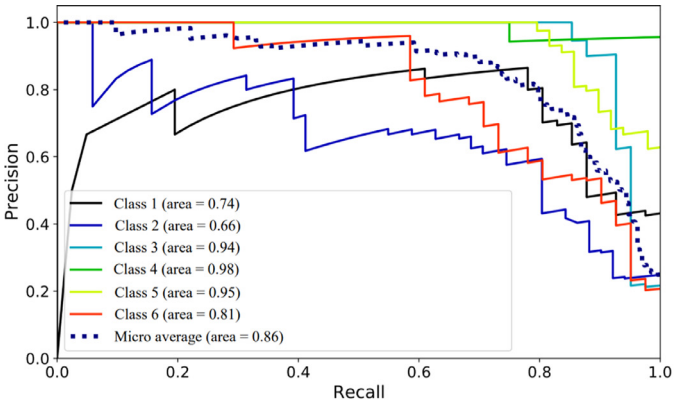
(b) ROC curve

Fig. 5. LSTM model classification result.

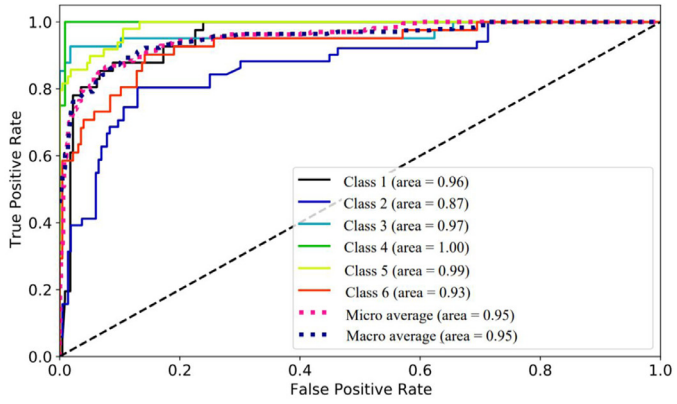
Table 1
Critical analysis of the methods.

Paper	Data-set	Method	Machine learning	Unsupervised	Adaptive learning	Performance
[29]	Twitter	Multi-model Features	Multimodal Depressive Dictionary Learning	No	No	81
[30]	Online Forum	Feature based	Naïve Bayes, Maximum Entropy, and Decision Tree	No	No	54.5
[27]	Erisk 17/18 | Anx 18/19	Bag-of-Words/Word Embeddings	One class SVM and KNN	No	No	62, 56, 77, 6.8
[21]	Amazon Mechanical Turk	Synonyms	Attention network	Yes	No	88.8
[28]	Erisk 17/18	TF/Embedding	Machine/Deep learning	No	No	66/61
Proposed	Amazon Mechanical Turk	Word Embeddings	Hyper graph attention learning	Semi-Supervised	Yes	95

Table 1, presents methods that have used the detection of depression in the context of mental disorders using texts [21,27–30]. These methods test the technique on the different datasets (eRisk 2017/2018, Twitter and Reddit). However, the semi-supervised learning method remains a viable way to improve the model. This study shows that node and edge level hypergraphs can help in training methods and intends to increase the number of trainable cases through a semantic extension approach. The proposed model aims to minimize the data annotation overhead. Therefore, the method helps in generalizing the learning system. The semantic vectors are grouped according to the hypergraph information that results from the context in which they occur. The resulting

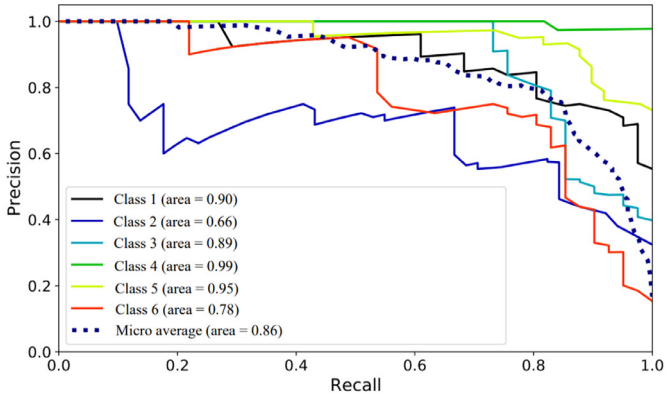


(a) Precision Recall Curve

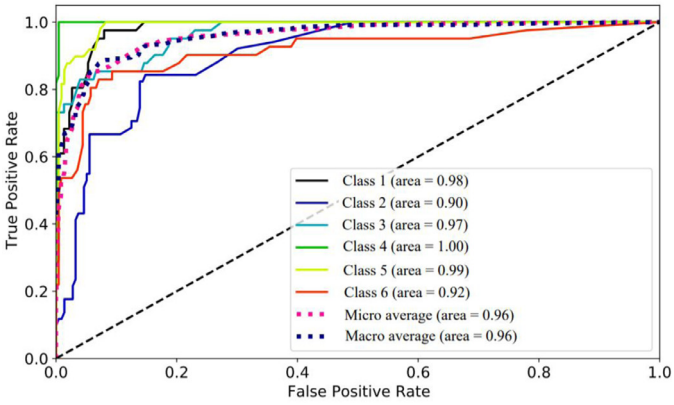


(b) ROC curve

Fig. 6. BILSTM model classification result.



(a) Precision Recall Curve



(b) ROC curve

Fig. 7. BILSTM with positional attention layer model classification result.

word embedding uses semantic information to select a subset of the unlabeled text. This method takes the unlabeled text from the hypergraph learning mechanism and labels the instances. The approach updates the model training with the new training points.

Together with the attention network, the stacked based on node and edge level model can extract critical work involving emotional significance and associations with the output class. The aggregation layers help maintain the vector representation of the informative terms of the nodes and edges. The process identifies emotional words to generate phrases that are then concatenated to make the document or conversation detectable for symptoms. When some words are deceptive and others are vital, the essential words are given more weight than their neighbors. The results show that adding additional vocabulary and grammatical variations can improve the performance of the model. The proposed method learns the meaning of sequences and applies attention-based weights to a vector representation that uses node-level and edge-level features. The representation is then aggregated with other words to produce sentence vectors. This aggregated vector contains all the semantic meaning and pattern information needed to categorize the instance class. The proposed system achieved a weighted F1 score of 0.95, and synonym expansion increased the training accuracy.

In addition, the built sentimental vocabulary subsequently helps reduce overfitting and generalization. The hierarchical attention-based model combined with the node- and edge-based hypergraph helps with vector representation. Node attention helps to enhance the emotionally triggering event, while edge attention enhances the context of the text. This tool can be used as an online adaptive intervention during virtual meetings with the psychiatric patient, with a psychiatrist providing the tool for reference notes.

5. Discussion

The proposed model must be run over longer epochs to obtain robustness results. However, it is worth noting that the linked features set reduces performance as training time increases. All techniques perform well on the classification problem when treated as a set of lexicons. When integrated with the attention network, the stacked-based on node and edge level model may extract key work with emotional relevance and links to the output class. The aggregation layers help keep the node's vector representation and informative edge words intact. The method decomposes inspirational words into phrases that are concatenated to find symptoms in a text or conversation.

Graph-based approaches outperform the others in terms of performance. This result suggests that capturing word interactions over long distances can directly improve text categorization ability. Sequence-based approaches (LSTM) outperform most graph-based baselines in classification. One possible explanation is that sequential context information is important in sentiment classification, but is not explicitly captured by the majority of existing graph-based algorithms. In particular, our model illustrates the value of high-order context information in learning word representations.

6. Conclusion

In this paper, we applied structural data to a sequencing challenge. We used the structure hypergraph, where nodes explore the structure of neighbors, which is useful for the attention-based LSTM model. We also embedded the model in an emotional lexicon to obtain a high-level sequence of nodes. Then, the weighted attention network is supported by the structural embedding. With an F-measure of 0.85, the experimental data show that the model

performs better than other models. In the future, we will apply automatic embedding selection before training to improve the selection of the language lexicon. In addition, the scalability of the model will be assessed using a similarity learning network. This will help to develop a more robust model for the correctness, reliability and integrity of the data.

Declaration of Competing Interest

Authors wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Authors confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed.

Authors further confirm that the order of authors listed in the manuscript has been approved by all of them.

Authors confirm that they have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing they authors that they have followed the regulations of our institutions concerning intellectual property.

Authors understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. Authors confirm that they have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email: jerrylin@ieee.org.

References

[1] S.L. James, D. Abate, K.H. Abate, S.M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017, Lancet 392 (10159) (2018) 1789–1858.

[2] M.G. Mazza, R.D. Lorenzo, C. Conte, S. Poletti, B. Vai, I. Bollettini, E.M.T. Melloni, R. Furlan, F. Ciceri, P. Rovere-Querini, F. Benedetti, Anxiety and depression in COVID-19 survivors: role of inflammatory and clinical predictors, Brain Behav. Immun. 89 (2020) 594–600.

[3] S.K. Mukhiya, J.D. Wake, Y. Inal, K.I. Pun, Y. Lamo, Adaptive elements in internet-delivered psychological treatment systems: systematic review, J. Med. Internet Res. 22 (11) (2020) e21066.

[4] S.K. Mukhiya, J.D. Wake, Y. Inal, Y. Lamo, Adaptive systems for internet-delivered psychological treatments, IEEE Access 8 (2020) 112220–112236.

[5] S.K. Mukhiya, U. Ahmed, F. Rabbi, K.I. Pun, Y. Lamo, Adaptation of idpt system based on patient-authored text data using nlp, in: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE, 2020, pp. 226–232.

[6] A. Konrad, V. Bellotti, N. Crenshaw, S. Tucker, L. Nelson, H. Du, P. Pirolli, S. Whittaker, Finding the adaptive sweet spot, in: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, 2015, pp. 17–21.

[7] E.A. Troyer, J.N. Kohn, S. Hong, Are we facing a crashing wave of neuropsychiatric sequelae of COVID-19? neuropsychiatric symptoms and potential immunologic mechanisms, Brain Behav. Immun. 87 (2020) 34–39.

[8] C. Karmen, R.C. Hsiung, T. Wetter, Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods, Comput. Methods Programs Biomed. 120 (1) (2015) 27–36.

[9] D.M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, S.S. Ghosh, Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: observational study, J. Med. Internet Res. 22 (10) (2020) e22635.

[10] J. Mühleck, S. Borse, E. Wunderer, B. Strauß, U. Berger, Online-befragung zur bekanntheit von angeboten zur aufklärung, prävention, beratung und nachsorge bei essstörungen, Prävent. Gesundheitsförderung 15 (1) (2019) 73–79.

[11] A. Neuraz, I. Lerner, W. Digan, N. Paris, R. Tsopra, A. Rogier, D. Baudoin, K.B. Cohen, A. Burgun, N. Garcelon, et al., Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the covid-19 pandemic, J. Med. Internet Res. 22 (8) (2020) e20773.

[12] E. Chen, K. Lerman, E. Ferrara, Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus twitter data set, JMIR Public Health Surveill. 6 (2) (2020) e19273.

[13] G. Nguyen, S. Dlugolinsky, M. Bobák, V.D. Tran, Á.L. García, I. Heredia, P. Malík, L. Hluchý, Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey, Artif. Intell. Rev. 52 (1) (2019) 77–124.

[14] T. Han, S. Niu, P. Wang, Multimodal-adaptive hierarchical network for multimedia sequential recommendation, Pattern Recognit. Lett. 152 (2021) 10–17.

[15] C. Ding, Y. Chen, Z. Liu, T. Liu, Prediction on transmission trajectory of COVID-19 based on particle swarm algorithm, Pattern Recognit. Lett. 152 (2021) 70–78.

[16] G. Siva Shankar, K. Manikandan, Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization, Pattern Recognit. Lett. 125 (2019) 432–438.

[17] V. Sze, Y.H. Chen, T.J. Yang, J.S. Emer, Efficient processing of deep neural networks: a tutorial and survey, Proc. IEEE 105 (12) (2017) 2295–2329.

[18] M. Wainberg, D. Merico, A. Delong, B.J. Frey, Deep learning in biomedicine, Nat. Biotechnol. 36 (9) (2018) 829–838.

[19] H.I. Fawaz, Deep learning for time series classification, CoRR (2020) arXiv:2010.00567.

[20] K. Kroenke, R.L. Spitzer, J.B. Williams, The PHQ-9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (9) (2001) 606–613.

[21] U. Ahmed, S.K. Mukhiya, G. Srivastava, Y. Lamo, J.C.-W. Lin, Attention-based deep entropy active learning using lexical algorithm for mental health treatment, Front. Psychol. 12 (2021) 471.

[22] K. Kroenke, R.L. Spitzer, J.B. Williams, The phq-15: validity of a new measure for evaluating the severity of somatic symptoms, Psychosom. Med. 64 (2) (2002) 258–266.

[23] W.H. Organization, et al., The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research, volume 2, 1993.

[24] L.F. Ribeiro, P.H. Saverese, D.R. Figueiredo, struc2vec: Learning node representations from structural identity, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017, pp. 385–394.

[25] S. Pan, P. Li, C. Yi, D. Zeng, Y.-C. Liang, G. Hu, Edge intelligence empowered urban traffic monitoring: a network tomography perspective, IEEE Trans. Intell. Transp. Syst. 22 (4) (2020) 2198–2211.

[26] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: The Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1480–1489.

[27] J. Aguilera, D.I.H. Farías, R.M. Ortega-Mendoza, M. Montes-y Gómez, Depression and anorexia detection in social media as a one-class classification problem, Appl. Intell. (2021) 1–16.

[28] R.M. Ortega-Mendoza, D.I. Hernández-Farías, M. Montes-y Gómez, L. Villaseñor-Pineda, Revealing traces of depression through personal statements analysis in social media, Artif Intell Med 123 (2022) 102202.

[29] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: a multimodal dictionary learning solution, in: IJCAI, 2017, pp. 3838–3844.

[30] L. Xu, R. Jin, F. Huang, Y. Zhou, Z. Li, M. Zhang, Development of computerized adaptive testing for emotion regulation, Front. Psychol. 11 (2020) 3340.