Expert Systems With Applications 182 (2021) 115265

Contents lists available at ScienceDirect

Expert Systems With Applications

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

# Negative emotions detection on online mental-health related patients texts using the deep learning with MHA-BCNN model

Kodati Dheeraj *, Tene Ramakrishnudu

*Department of Computer Science and Engineering, National Institute of Technology, Warangal, Telangana 506004, India*

ARTICLE INFO

*Keywords:*
Emotion analysis
Mental-health-care
GloVe-embeddings
Deep learning
MHA-BCNN

ABSTRACT

Mining the emotions in the text related to mental health-care oriented is a challenging aspect, especially dealing with a long-text sequence of data. The extraction of emotions depends upon the various psychological depression factors like negative and ambiguity. Identifying these factors is the most perplexing task for every psychiatrist to treat their patients. Our study includes the deep learning (DL) models with global vector representations (GloVe) embeddings to capture the text sequence of data. We proposed a model multi-head attention with bidirectional long short-term memory and convolutional neural network (MHA-BCNN) is a pre-eminent mechanism that outperforms better than past research works for capturing the negative text-based emotions. In this paper, by using DL extracted the various negative mental-health emotions like addiction, anxiety, depression, insomnia, stress, and obsessive cleaning disorder (OCD). By using the GloVe embeddings and handled the ambiguity factors like multiple emotion words in a certain sequence. As we proposed a vigorous appliance in our research to capture and hoard the long-term dependencies. We extracted the questions related to mental health issues were posted by the patients in an online mental healthcare-oriented platform. We efficaciously handled both negative and ambiguity factors at the document level. Our suggested exemplary MHA-BCNN surmounts various aspects from preceding research works and ensued preeminent performance. Experimental results show that our proposed framework MHA-BCNN outperformed than the erstwhile research works.

## 1. Introduction

Identifying the emotions in a text sequence is generally considered to be a difficult task, the main reason is long-term dependencies are found in a text. Recently, there has been a huge development in the field of emotion analysis, but studies regarding text sequence is a vital chore to know about the emotion patterns are generated by the sequential text. Many emotion analysis frameworks have been proposed in the previous literature. Healthcare-related online amenities contain various text-related requests allied with depressions. Both psychiatrists, mental-health specialists are reviewing and responding to these requests and need to address them aptly to signify the specific type of depression indicated. The online mental-health care platform is an opulent source of textual content reflecting how individuals recognize different facets of their life. Text emotions play a substantial part to recognise mental-health situations. Our research recommends that text expression identification is a hopeful path of inquiry in mental healthcare, the text expressions can also reflect the individual's mental state. Addressing from the erstwhile studies about the online mental-health platform as social conduct with positive or negative effects and results have been abstracted and measured from patients texts. The current study was directed to find out the associations of mental health through emotional intelligence and behavior factors.

This study inspects the conduits through which mood discernment affects, stress, and symptom reporting among various patients over mental-health care issues (Al-Abri & Al-Balushi, 2014). It examines the positive effect, and the negative effect (Mekruksavanich, Jitpattanakul, & Hnoohom, 2020). In psychological healthcare-related online service (Bai, Lin, Chen, & LIU, 2001), the text-based questions from the various patients are important to analyze the depression level. In general, after reading their question on online health care, both psychiatrists and mental-health specialists usually give their suggestions to their patients. Continuous emotion recognition is a type of document level representation, where extracting the emotions in each document (Oh, Lee, Ko, & Choi, 2017; Rhanoui, Mikram, Yousfi, & Barzali, 2019). Identifying the mental health emotions in the online platform by DL is the most challenging aspect (Wongkoblap, Vadillo, & Curcin, 2017). By using DL models like neural networks (NN) to tenacity the problem of multi-label

* Corresponding author.
*E-mail addresses:* dheerajkodaticse@student.nitw.ac.in (K. Dheeraj), trk@nitw.ac.in (T. Ramakrishnudu).

https://doi.org/10.1016/j.eswa.2021.115265
Received 15 March 2021; Received in revised form 19 May 2021; Accepted 20 May 2021
Available online 30 May 2021

classification. The word embeddings and convolutional neural networks (CNN), and long short-term memory (LSTM) are used to develop the prevailing classifier to recognize emotion labels. These models have been effectively applied in a broad range of classification tasks are used here to build a vigorous classifying contrivance for the multiple emotion labels within mental-health texts based on the classification performance. Understanding the emotion recognition in text sequence is a multidisciplinary task involving various factors like attention models (Luong, Pham, & Manning, 2015) and long-term dependencies, to capture the mental-health texts.

Sailunaz and Alhajj (2019) detected the basic emotions with sentiments like "joy, fear, anger, sadness, surprise, disgust and neutral" on text documents. However, by considering the basic emotions with "joy" and "neutral sentiments", which do not impact the mental-health state. Similarly, Chatterjee et al. (2018) proposed a deep learning approach for detecting the text emotions which include "happy, sad, and angry" and mentioned that the emotions are related to the physiological states. Xu et al. (2020) detection with microblog texts, where the approach related to emotion detection rather than capturing text-based emotions with semantic features and text sequence. Wu, He, Yu, and Lai (2020) worked on text-based mental-health emotion labels detection for only short text documents. The above works motivated us to analyze as follows:

- Considering the mental health emotion detection which affects the individual's behavior and maintain the long-distance semantic information.
- By taking the word embeddings to capture semantic features with the text sequence and attention on each word.
- Further consideration like capturing the mental health emotions with long-text documents rather than short-texts.
- Also, including the multiple emotion categories with text sequence, can identify the impact on mental health rather than fewer emotions.

When dealing with emotion analysis for mental health can be divided into various psychological factors (1) positive psychological factors (2) negative psychological factors (3) ambiguous psychological factors. In case of positive psychological factors can be further categorized into various types like happy, excitement, cheerful, laughable, smile, etc. (Al-Abri & Al-Balushi, 2014; Zgierska, Rabago, & Miller, 2014). Whereas in the negative psychological factors like addiction, anxiety, depression, insomnia, stress, and OCD (Wu et al., 2020). The ambiguous psychological factors deal with ambiguous emotions and uncertainty of text, for example, "Expert advice needed for anger and depression for saving the marriage. I need expert advice on my wife's behavior. Regarding my wife's issue, she gets angry frequently without reason. When she gets angry she shouts and even try to hit me few times and remains angry this way for few hours my family tries to avoid talks with her thinking she might get angry for some or other reasons and blame them. She is unable to mix with people and gets afraid or even depressed. She gets angry when she is tired of the work and when she is unable to find any solution". This sentence is related to the mental health issue, here the underlined word represents the multiple emotions. These factors are situation-oriented, although this type of mental-health posts are not having one obvious emotion which is situationally dependent.

So, need to capture these emotions with the situational changes in the user's text, which are posted to the psychiatrist by using LSTM-attention techniques combining with other deep neural network techniques. LSTM is a type of RNN that solves "long-term dependencies", also another type of LSTM like BiLSTM stands for " bidirectional LSTM" which solves the unidirectional LSTM captures only previous information whereas BiLSTM captures both previous information and predicts the future information, as it moves both the forward and backward direction. So the problem here, when dealing with BiLSTM in a text sequence information, is to capture the text which is long-term dependent. So, this paper included attention techniques like self-attention (SA) and multi-head attention (MHA) techniques.

The SA technique can capture better only for short-length text documents, but in MHA can capture the information which has a large amount of data. Dealing with negative psychological factors and ambiguous psychological factors is a stimulating chore. The proposed work is focused on the multi-labeled classification problem, the text contains multiple emotion labels as provided the information in the ambiguous psychological factors. Here, manually labeled the data using the mental health care oriented lexicons and considered as target (Wu et al., 2020). To find emotions in large text sequence data, by GloVe embeddings (Pennington, Socher, & Manning, 2014), using the cosine similarity to store the inputs into a vectorized format.

Table 1 represents words category concerning the emotion label and their weights representing various hidden vectors, for instance, the first-word category can be considered as drug, medication, withdrawal, sedative, dose these are the expedient features to detect the emotion label. As emotions are manually labeled using the word category, these terms are related to the psychiatric emotion dictionary and verified by clinicians. By training DL models using mental health-related questions with manual labelings as our target values. Following this, from the GloVe embeddings to find the similarity between the two vectors, where our words are stored in a vectorized format, moreover considering the word-embeddings instead of taking the bag-of-words (BoW) under Section 3. Using the LSTM-CNN features can compete for the models like SVM for dealing with short text documents (Chen, Lv, Wang, Li, & Wang, 2019), whereas the continuous large sequence text as an input, can consider the GloVe embeddings (Pennington et al., 2014).

DL is the best approach and becoming popular for large-sequence of textual data when dealing with the emotions (Zhang, Wang, & Liu, 2018), text with emotion expression also frolicked an imperative part to implement the additional attention to this area (Seyeditabari, Tabari, & Zadrozny, 2018). The response-related emotion finding models consist of two parts, extract and generation part these are used on the input utterance using the LSTM model (Choi & Lee, 2020). Working on a long text sequence with multi-labeled classification (Chen, Ye, Xing, Chen, & Cambria, 2017) used only the CNN-RNN model to extract the label correlations. With mental-health questions as our model trainings and labels consider as target values and magnificently extracted the negative psychological emotions. Achieved better validation accuracy using MHA-BCNN with GloVe embeddings which outperforms than erstwhile research works. This paper provides experiments using DL models with larger data from a psychiatrist text. In section 1, a brief introduction about the mental health care texts, section 2 about previous research works, section 3 about word embeddings and DL models, section 4 model representation, section 5 about experimental results, section 6 conclusion and future scope.

## 2. Literature review

Emotions broadly affect the decision-making of individuals, this computing takes emotion circumstances into account intending to modify decision support to individuals. Nevertheless, the accurate identification of emotions within the novel text represents a stimulating task due to the intricacy and vagueness of language. Although DL has progressed as the advanced in many tasks for text mining, its benefits

**Table 1**
Word lexicons for the emotion labels.

| Emotion Labels | Word Lexicons |
|---|---|
| **Addiction** | drug, medication, withdrawal, sedative, dose. |
| **Anxiety** | stiffness, crowd, ridicule, shivering, tense, exited. |
| **Depression** | disorder, depressed, sad, mood, suicide, disturbance, anger. |
| **Insomnia** | nightmare, sleep, insomnia, dream, awake. |
| **OCD** | OCD, thoughts, behavior, compulsion, personality disorder, overeating, depersonalization. |
| **Stress** | strian, nervous, suffering, upset, weight. |

concerning affective computing are not yet understood (Kratzwald, Ilic, Kraus, Feuerriegel, & Prendinger, 2018). Balahur, Hermida, and Montoyo (2012) proposed the seven types of emotions like “anger, disgust, fear, guilt, joy, sadness, and shame” are the most commonly used paradigm in psychology. Our literature survey can be additionally classified into two aspects (1) psychological-healthcare emotion recognition (2) emotion recognition with DL models.

### 2.1. Psychological-healthcare emotion recognition

To extract emotions in the information sequence is important, extracting the emotions in this scenario depends on the long-term dependence of data. The online health group for patient-doctor communication is the furthermost substantial aspect. A large combination of text data is generated during the patient-doctor communication process. It is primarily explored the impacts of online patient-doctor communication content on patient questions. Mental-healthcare emotions are patient-centered, considering a large number of mental health texts. The prominence of calculating the emotions and enlightening the results that are related to patients texts. Including patients requirements and outlooks are measured as mandatory for eminence expansion in the healthcare system (Al-Abri & Al-Balushi, 2014; Zgierska et al., 2014). Wu et al. (2020) proposed the BiLSTM-CNN model, for identifying the emotions in psychiatric data and used Chinese text corpus to train word embeddings, to achieve the most common semantic relation for each word rather than to capture the emotions in long text documents. Bahgat, Wilson, and Magdy (2020) adopted a novel approach for analyzing mental health-related content by using word-embeddings, handicraft lexicons, and psycholinguistic approaches to improve classification accuracy, but only used smaller corpus size subreddits.

Fei et al. (2020) used a Deep CNN over mental-health care, but they need to mention the emotion category changes with the situation. Xia, Zhao, Wang, Liu, and Ma (2018) proposed the LSTM model to extract the emotions from Chinese patients comments and they used the skip-gram model for word embedding and they got 89% F1-score using LSTM against the five ML methods namely SVM, Random Forest, Decision Trees, Gradient Boost but not mentioned about the emotion categories. Chen, Guo, Wu, and Ju (2020) worked on unstructured data from the online patient-doctor interaction community and used text data mining and econometric analysis models but not mentioned sequence data. Hipson (2019) taken children’s and adolescent’s online published poems from grade 4 to grade 12, using polynomial regressions and linear, non-linear trends and demonstrated results as sinusoidal patterns but not mentioned much about emotion data between children’s vs adolescent’s age group. Ireland, Hassanzadeh, and Tran (2018) used conversation chatbot for sentiment analysis to extract negative and positive utterances on E-health, but need to mention about the emotion analysis. Oh et al. (2017) worked for mental-healthcare related continuous emotion identification, they used clinical psychiatric counseling data for multi-model using the natural language understanding (NLU).

Cox, Moscardini, Cohen, and Tucker (2020) worked on machine learning (ML) to address the basic questions on suicide at several stages of examination. Brailovskaia and Margraf (2016) collected the data from the group of both the Facebook users and non-Facebook users, observed that the non-Facebook users having fewer depression levels than the Facebook users, this is because of the high level of neuroticism. The mental ailments embrace various disorders likewise, depression is the utmost noticeable. Furthermore, both anxiety and depression syndromes can cause suicidal risk. Wongkoblap et al. (2017) presented a detailed comparison of the approaches proposed on the mental health disorders connected with social network data. Alm, Roth, and Sproat (2005) proposed the SNoW learning for text-based emotion extraction, they used the BoW for content based on the classification of emotions with non-emotions by considering basic emotions from large dataset like anger, fear, sadness, etc., Kraiss, ten Klooster, Moskowitz, and Bohlmeijer (2020) proposed meta-analyze suggestion concerning the rapport amongst emotion recognition on quantifiable samples. Pânisoara, Panisoara, and Sandu (2015) worked on analyzing the psychological emotions of the students by gender. Huang et al. (2015) proposed a topic model for categorizing suicidal risk sway in Chinese microblog.

### 2.2. Emotion recognition with DL models

The DL models provide prevailing erudition contrivances for various categories of applications. The deep neural networks (DNN) are multiple feed-forward neural networks and normally practice the approximation theorem to estimate any multifaceted constant gatherings specified by sufficient neurons to develop the NN model. Kratzwald et al. (2018) proposed sent2affect is a customized form of transfer learning used for affective computing, as it was used in pre-trained BiLSTM model outperformed than traditional ML models with BiLSTM the F1-score about 68.8%. Almeida, Cerri, Paraiso, Mantovani, and Junior (2018) used an ensemble model to detect six types of emotions in multi-label classification for only short texts and achieved about 64.89% F1-score. Balahur et al. (2012) used EmotiNet with the comparison of the individual evaluation like appraisal criteria with well-established methods. Partaourides, Papadamou, Kourtellis, Leontiades, and Chatzis (2020) considered some approaches for online social network (OSN) analysis but proposed only self-attentive-based text sequence recognition but not mentioned the long-text sequence of data. Zhang et al. (2020) used an LSTM dyadic conversation but not mentioned much about emotion identification. Habimana, Li, Li, gu, and Yu (2020) proposed generative adversarial networks and bi-direction encoder for multi-lingual features by using DL models and need to mention both dynamic heterogeneity information. Poornima and Priya (2020) not mentioned the text with multiple emotion categories have been considered.

Song, Zhao, Liu, and Zhao (2018) for extracting emotions from Webio texts using CNN with BiLSTM and got an accuracy rate of about 90.2 % which is 2.5% better than BiLSTM and 1% better than LSTM with CNN, but not mentioned about emotion categories. Sun and He (2020) proposed a hybrid neural network model for multi-granularity data on short texts, which outperformed against the many single neural network methods. Lee, Kim, and Song (2020) used a semi-supervised model where small amounts of labeled data and performed to the large datasets. Li, Li, Sun, and Wang (2017) worked only on RNN with short text documents rather than the emotions where positive and negative sentiments were categorized using the mobile phone reviews dataset. Li, Feng, Wang, and Zhang (2019) proposed CNN and LSTM with MHA based neural network model for identifying the context clause, candidate clause, and emotion clause, this model outperformed previous models but when considering backward LSTM, it needs to calculate the loss function at each layer may result with computational overhead when considering this scenario and also need to capture more than one emotion word. Lin, Li, Yang, Xu, and Lin (2020) used BiLSTM, CE-B-MHA first generates word vectors based on text and used BiLSTM. It can capture the context relationship encode word sequences initially, which results that overly complex models are prone to serious overfitting and the simpler models get better serious. When the sample size is large, the model performance decreases even LSTM cells can’t capture long terms dependencies, they start to perform lesser as the sequence length increases.

Lin, Su, Chien, Tsai, and Wang (2020) SA model that initiates a weight matrix, which is used to coalesce the hidden states from the two LSTM models with matrix multiplication. Computational overhead by exploring various neural network-based architecture (RNN and LSTM) models for performing the sentimental analysis chore on the text data corpus, but not performed well on the multi-labeled classification. For the bidirectional LSTM, considering an embedding layer and instead of loading random weights, allowing to load the weights from the GloVe embeddings. Brueckner and Schulter (2014) used BiLSTM for sentiment analysis chore, which contributes to the elucidation of attaining both erstwhile information and imminent information by using the bidirectional mechanism, but not mentioned capturing long term dependencies, rely on single-layer Bidirectional RNNs. Knowledge (2020) used API plus service specification like NLP interchange to senpy open-source service for framework evaluation between developers and users, this solves the heterogeneity issues and outcomes of the emotion recognition services but only got 0.607% accuracy from a series of public datasets. Sathis Kumar, Mohamed Nabeem, Manoj, and Jeyachandran (2020) not mentioned sequence identification in multi-labeled classification for emotion recognition, only mentioned the SVM model to capture the current emotions.

Vanlalawmpuia and Lalhmingliana (2020) used Twitter tweets by SVM model not captured long text sequences. With MHA named CMA-MemNet (Zhang, Xu, & Zhao, 2020) was improved over memory networks to capture information at word level and also proposed a convolution to capture n-gram grammatical information but need to perform semantic information at the sentence level. Barron Estrada, Zatarain Cabada, Oramas, and Graff (2020) proposed the algorithm model called EvoMSA for educational opinions classification which outperformed erstwhile models but only a few emotion labels were considered for the eduSERE data corpus and got about 84% accuracy. Boiy and Moens (2009) performed sentiment classification, and performance increased by filtering neutral sentence and using SVM they got better performance but not mentioned much about emotion recognition and they ignored sentence with more emotion categories. Koppel and Schler (2006) combined pairwise classifiers on neutral information by using optimal stack but performed only on binary classification. Ren, Xu, Lin, Liu, and Yang (2020) used the Twitter dataset to identify the high-level features for sarcasm expressions by using local max-pooling over multi-level memory but they need to mention much about dyadic communication.

Makki, Alhalabi, and Adham (2019) proposed HF-EMA to identify the virtual reality over wearable devices in Twitter user tweets by supervised learning model in terms of ROC and accuracy but not mentioned sequence information. Sailunaz and Alhajj (2019) worked on the ambiguity of text emotions on Twitter tweets and replies using the k-means clustering model and not mentioned much about sequential information. Choi and Lee (2020) worked on the LSTM model for feature extraction using the pre-trained values, and Devlin, Chang, Lee, and Toutanova (2019) BERT model is a fine-tuned model that represents only encoder attention. Batbaatar, Li, and Ryu (2019) worked on semantic emotion detection using the SENN model by taking GRU with pre-trained word embeddings. Rhanoui et al. (2019) proposed CNN with BiLSTM for the document-level depiction to ascertain the large sequence texts.

### 2.3. Considerations over existing works

The model MHA-BCNN with Glove embeddings maintains the semantic relations between the words rather than the CMA-MemNet model (Zhang et al., 2020) only preserves the information randomly rather than maintaining the semantic relations. With our proposed model, distributing the pre-trained embeddings with two independent LSTMs unlike (Xia et al., 2018) LSTM with a skip-gram model for only a single LSTM preserves the information from the past. Similarly, with the LSTM-CNN model (Chen et al., 2019) and LSTM (Li et al., 2019), CNN and LSTM with MHA model, these models store the information in unidirectional LSTM. Our model stores the information in both forward and backward LSTMs to hold the long-term dependencies. Batbaatar et al. (2019) worked on semantic emotion detection using the SENN model by taking GRU with pre-trained word embeddings, difficult with GRU model providing less consideration for information preserving by taking the fewer parameters and contains slow convergence and less efficiency. The proposed model resolved the issues from GRU by spreading the information for independent LSTMs and providing semantic relations with text sequences. Our model resolved the long-term dependencies problem occurring during the CNN-RNN model (Chen et al., 2017), RNN model for short text documents (Li et al., 2017). Where the RNN model suffers from long-term dependency.

The BiLSTM (Almeida et al., 2018) for short texts and (Wu et al., 2020; Song et al., 2018) proposed BiLSTM with CNN model for emotion detection, where these BiLSTM and BiLSTM with CNN models do not maintain the positional embeddings. These embeddings mention that the proposed model compares not only adjacent words in a certain sequence but also compares word which is far away from the particular word. As our model maintains the attention on each word rather than comparing with only adjacent words. Rhanoui et al. (2019) proposed CNN with BiLSTM for the document-level not maintaining semantic relation like the proposed model. Our model provides further attention to each ‘word’, it can correlate with the other words which are far away from the particular word within a sequence and performs better than BiLSTM (Almeida et al., 2018), BiLSTM-CNN (Wu et al., 2020), and GRU (Batbaatar et al., 2019) models. Our model maintains multiple heads associated with the linear layer, where each head represents the position of the input features so that providing further attention to each token within the certain sequence unlike Lin et al. (2020) and Ambartsoumian and Popowich (2018) worked on self-attention providing the single attention mechanism within the sentence. Gao et al. (2018) considered an attention-based BiLSTM network for emotion classification but the model performance decreases when considering long-text documents. Another model TextCNN with TF-IDF (Song, Geng, & Li, 2019) not capturing the word co-occurrences when comparing to our proposed model with Glove embeddings maintains the word co-occurrences. Also, Fesseha, Xiong, Emiru, Diallo, and Dahou (2021) worked on CBOW but unable to provide the semantic relations, and also they proposed word2vec with CNN not suitable for long-text documents.

Moreover comparing with other erstwhile works for emotion detection like (Kumar et al., 2020) worked on text-based emotion identification using the SVM. However, using SVM it is firmer to classify multiple emotion categories. Both (Wang et al., 2020) with SVM, logistic regression models for text classification, and (Rajabi, Shehu, & Uzuner, 2020) proposed multi-channel with CNN-BiLSTM model for multi-labeled emotion classification these works provide less performance for maintaining the semantic relations between the words. Many researchers proposed traditional ML techniques like logistic regression, SVM, random forests for detecting the emotions within the sentence. However, these techniques do not perform better with multiple emotion categories and losing the information when considering long-text documents. Our model works better with multiple emotion categories and solves the long-term dependency problem and maintains the semantic relations between the words and providing the attention of each word in a sequence. Also, our model includes the CNN layer for classification result generated in the output layer unlike (Xia et al., 2018) with the LSTM model holding the previous information. This paper shows the best results with the proposed model when compared to existing works for detecting text-based emotions and the performance of the proposed mechanism is shown in Section 5.

## 3. Preliminaries

### 3.1. Problems with BoW

The BoW is represented as a word-count vector for each document. In DL, considering the sparse matrix will be accompanied by an enormous sum of training data. In general, it is not sagacious to use the BoW for the data contains a long-text sequence, the major disadvantages like the vast amount of input vectors through a huge amount of weights for a general neural network. Next, these are computationally exhaustive, where the new weights related to addition computation, which are included to train and predict. Absence of evocative relations and without concern for semantic relation between the words. The BoW without considering the illustration of the sequence in which they emerge. The word embeddings is a procedure where distinct words of a document or domain are signified as actual-valued vectors in a lower-dimensional space (Poornima & Priya, 2020).

### 3.2. GloVe embeddings

Capturing emotions in long text sequence data is a vigorous task especially dealing with the mental health-care oriented field. The method proposed in this paper is focused on both negative psychological factors and ambiguous psychological factors using mental health texts. The labels from various psychiatrists and mental health-related questions are extracted manually using the lexicon of psychiatric terms. To train our DL model using the larger sequence of data was a challenging task. The word embeddings having a larger corpus can train using the GloVe method, also it can train domain corpus. The GloVe method enhances the co-occurrence likelihood of words 'i' and 'j', these will be executed based on the combined co-occurrence of words from the corpus (Pennington et al., 2014). $Sim(V_i, V_j,)$, calculating the similarity between the two vectors.

$$Cos\left(V_i, V_j\right) = \frac{V_i * V_j}{\|V_i\| * \|V_j\|} \tag{1}$$

In this research, implemented the GloVe embeddings to develop the word co-occurrence for the emotion category recognition.

### 3.3. Deep neural network models

#### 3.3.1. CNN

The CNN stands for convolution neural network is a prevailing model for feature extraction, which is exactly stimulating in the opinion analysis, particularly when related with long-text data, where the mining of its structures remains tough. The convolution layer purpose to discover the union between the various sentences or documents in the text, using filters of size 'f'. A CNN is a feature-mining model and is intended to be combined in a larger network (Goldberg, 2017). The n-gram is a subsequence of '*n*' contiguous words erected from the given sequence. The important standard is to calculate, from an assumed sequence of words '*w*' and the likelihood function of the advent of the subsequent word denotes as follows:

$$p(w_i|w_1, \ldots, w_i - 1) = p(w_i\|w_i - (n-1), w_i - (n-2), \ldots, w_i - 1) \tag{2}$$

#### 3.3.2. LSTM

LSTM stands for long short term memory is a kind of RNN connected with the previous information to the present task. LSTMs have an advantage comparing with CNN and RNN in several ways because of the selective retention of patterns for long extents of time. All RNN's contain the procedure of a chain of recapping the segments of a neural network (Xia et al., 2018). In regular RNN's, these recapping segments will have an eminently manageable edifice, likewise a single tanh layer. The LSTM calculates as follows:

$$C_t = f_t * C_{t-1} + i_t * \widetilde{c_t} \tag{3}$$

$$f_t = \sigma\left(W_t * \left[h_{t-1}, h_t\right] + b_f\right) \tag{4}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{5}$$

$$\widetilde{c_t} = tanh(W_c * [h_{t-1}, x_t] + b_c) \tag{6}$$

$$o_t = \sigma(W_i * [h_{t-1}, x_t] + b_o) \tag{7}$$

$$h_t = o_t * tanh(c_t) \tag{8}$$

Where $C_t$ is the candidate cell state, $C_{t-1}$ is the old cell state, $\widetilde{c_t}$ is the cell input activation vector, $x_t$ is the current input vector, $f_t$ is the forget cell, $i_t$ and $o_t$ are input and output gates, $h_t$ is the output vector of LSTM unit, $h_{t-1}$ previous hidden state, $W_t, W_i, W_c, W_i$ are weight matrices and $b_f$, $b_i$, $b_c, b_o$ are learnable parameters used for training, both the $\sigma$ and tanh are the activation functions correspondingly and a logistic sigmoid function, also '*' denotes the element-wise multiplication.

#### 3.3.3. BiLSTM

The BiLSTM stands for bidirectional long-short term memory, connects with two hidden layers of opposite directions. In this type of procreant DL, from the output layer can gather the information moving to the past i.e, backward and heading towards future, the forward states together. The Bidirectional LSTM model preserves two distinct states for both the forward and backward inputs that are produced by two dissimilar LSTMs. The first LSTM is a consistent sequence that appears from the starting of the document, whereas in the next LSTM, the sequence of input was passed in the opposite order. The idea related to a bidirectional network is to extract the information of contiguous inputs. Both forward and backward pass would be considered to calculate the output $\widehat{y}$ at a time 't' (Xie, Chen, Gu, Liang, & Xu, 2019; XinSheng & Teng, 2019). Considering the input as $x_t$ and $(h_t - 1)$ hidden state of forward LSTM can be calculated as:

$$\overrightarrow{h}_t = LSTM\left(x_t, \overrightarrow{h}_{t-1}\right) \tag{9}$$

Where as the backward LSTM can be calculated as.

$$\overleftarrow{h_t} = LSTM\left(x_t, \overleftarrow{h}_{t-1}\right) \tag{10}$$

From the concatenated output of both forward LSTM and the backward LSTM.

$$h_t = \left(\overrightarrow{h}_{t-1}, \overleftarrow{h}_{t-1}\right) \tag{11}$$

#### 3.3.4. GRU

The GRU, well-known as the gated recurrent unit is a kind of RNN architecture, where GRU is analogous to LSTM components. The GRU includes both reset and update gates in lieu of LSTMs forget, input and output gates. The reset gate regulates how to coalesce the current input through the preceding memory, whereas the update gate outlines in what way copious of the preceding memory to preserve around (Batbaatar et al., 2019). In GRU, considering the reset gate to all 1's and the update gate to all 0's which representing again at our general RNN model. For the GRU the hidden state $h_t$ can be computed as follows:

$$z_t = \sigma\left(W^{(Z)} i_t + U^{(Z)}\ h_{t-1}\right) \tag{12}$$

$$r_t = \sigma\left(W^{(r)} i_t + U^{(r)} h_{t-1}\right) \tag{13}$$

$$\hat{h}_t = tanh\left(Wi_t + r_t * Uh_{t-1}\right) \tag{14}$$

$$h_t = (1 - Z_t) *\ h_{t-1} + Z_t * \hat{h}_t \tag{15}$$

Where $z_t, r_t$ denotes are the reset and update gate vectors, $h_t, \hat{h}_t$ are output and candidate vectors, $\sigma$ is the sigmoid function.

## 4. Methodology

### 4.1. Text mining with attention mechanisms

To capture the emotions in psychiatric text, to reformulate the classification chore as a collating problem, and to propose a comparison to enhance BiLSTM. Analyzing the fluctuation in the performance of emotion analysis when multiple features are considered. Instead of considering the self-attention mechanism, to use the MHA mechanism especially, working on a long-text sequence. In this research, mentioned previous research works on both the sequential information and mental-health text on emotion analysis using DL techniques. Also, addressed the gaps from both psychiatric text and sequential information on emotion analysis from erstwhile research works. Our goal is to find the emotion analysis when dealing with multi-labeled classification using the BiLSTM plus MHA on a psychiatric text-related patient-doctor conversation. Moreover, by taking patient questions as input and it is required to identify any situational changes that happened in the middle it is essential to capture those changes. Also, to optimize the loss function when applying DL methods. Previous researchers proposed BiLSTM for emotions in sequential information with CNN using psychiatric texts, although to compute our loss function at each layer by applying the back-propagation, this results in the time-consuming task when moving back to the initial stage. Instead of this, our model MHA-BCNN is recommendable to minimize our residuals without any computational overhead.

#### 4.1.1. General attention mechanism

The attention mechanism is illustrated as plotting a query from previous layer hidden features, which includes a combination of key-value sets to output, where the vectors related to the query, keys, and values, output. Moreover, the weighted sum of the values is calculated from the output, whereas the weight apportioned to every value is calculated by the similarity representation of a query by the equivalent key. Bahdanau, Cho, and Bengio (2014) used the BiLSTM, it generates the sequence of annotations $(h_1, h_2, \dots, hTx)$ for every input document.

$$h_j = \left[\overrightarrow{h}_j^T, \overleftarrow{h}_j^T\right]^T \tag{16}$$

Altogether the vectors indicate $h_1, h_2, \dots, h_n$ used in Bahdanau et al. (2014) are fundamentally the sequence of both hidden states in the encoder like forward and backward states. In general, considering the vectors $[h_1, h_2, h_3 \dots, hTx]$ are the proceedings of Tx, amount of words from the input documents. In the general encoder-decoder architecture, encoder LSTMs final state can be used (hTx in this case) as the context vector. The attention model includes queries as 'q' and a combination of key-value pairs can be represented as (k, v), attention model can be comprehensive to calculate a weighted amount from the related values on the 'q', also the equivalent keys. The 'q' regulates related values to emphasize and the query can generalize as the 'q' attends to the values. Moreover, assuming the earlier decoder position related to the query vector, also the hidden states of the encoder as 'k' and value vectors, it indicated as the output of the weighted average, whereas the weights are related by the similarity function between the 'q' query and 'k' keys. A reminder is both keys and values can be various gatherings of vectors (Vaswani et al., 2017), so the key and value considering the same amount of features. The query 'q' values $(v_1, .., v_n)$ and keys $(k_1, .., k_n)$.

To calculate output as $o_i$ and $y_i$ is formed by the weighted amount of annotations.

$$o_i = \Sigma_{j=1}^{n} a_j(v_j) \tag{17}$$

The weights $a_j$ calculated by the softmax function and equation results as follows.

$$a_{ij} = \frac{softmax(v_j) = exp\, f(k_j, q)}{\Sigma_{i=1}^{n}\ exp\, f(k_i, q)} \tag{18}$$

For the similarity function 'f', assuming the scaled dot-product attention from Vaswani et al. (2017).

$$f\left(k, q\right) = \frac{q * k^T}{\sqrt{d_k}} \tag{19}$$

Here, $d_k$ denotes the dimensionality of key vectors. The scaling is completed to progress statistical solidity of the queries 'q' and $d_k$.

#### 4.1.2. SA mechanism

The essential process of every transformer model is the SA process. The SA is a sequence-to-sequence process, also considers the input as a sequence of vectors, so the resulted output generates as the sequence of vectors, where the SA layer outputs are served to the feed-forward neural network (FNN). SA, occasionally known as intra-attention, is the attention method indicating various locations of a particular sequence in direction to calculate the illustration of a sequence (Vaswani et al., 2017). Corresponding feed-forward network is independently applied to each position. Considering the sequence input as S= $(x_1, x_2, \dots, x_n)$ of words is transformed into a sequence $Y = (y_1, y_2, \dots, y_n)$. The dimension $d_k$ for all vectors, to generate the vector $y_i$ from the output, the self-attention mechanism aptly considers the weighted average sum over to every input vector.

$$y_i = \Sigma_{i=1}^{n}\ w_{ij} x_j \tag{20}$$

Here the 'j' indicates the indexes to the entire sequence, whereas the '$w$' is related to the weighted amount for the entire 'j'. Although $w_{ij}$ weight is not a constraint, as in an ordinary neural network, so the $x_i$ and $x_j$ are the resultant functions. The dot product is the meekest choice for this function.

$$w_{ij} = x_i^T x_j \tag{21}$$

Here, the input vector $x_i$ from the same location related the present $y_i$ the output vector, receiving a completely new sequence of dot products for the next output sequence vector, with a distinct weighted sum. The dot product provides resulted amongst positive infinity or negative, also implementing a softmax function [0,1] to map the values and to confirm that, the values are summation to '1' to the entire sequence can be defined as follows:

$$w_{ij} = \frac{exp\ w_{ij}}{\Sigma_{i=1}^{n}\ w_{ij}} \tag{22}$$

To enumerate alignment using basic dot-product attention, the set of comparisons used to determine context vectors can be reduced as follows.

$$Attention\left(q, k, v\right) = \frac{softmax\left(q * k^T\right)}{\sqrt{dk}} * v \tag{23}$$

#### 4.1.3. MHA-BCNN model

MHA is the extension of SA and the MHA mechanism allows the mechanism for combined attention to inputs from various illustration subspaces at various locations. MHA combines the various attention contrivances to make complete attention towards text documents. The MHA mechanism arranges the words related to the sequence by the words in the sequence, thus computing the demonstration of a sequence. MHA mechanism is represented in (Eq. (24)), here the query and key, the value represents through a linear layer transformation and then arrives into the scaled dot-product layer Fig. 1.

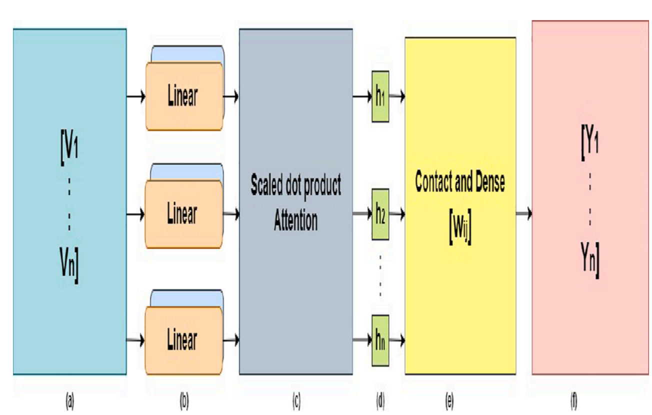



**Fig. 1.** The neural-architecture of a MHA model. (a) input features (b) linear layers with multiple-heads (c) scaled dot-product (d) attention-head features (e) normalize (f) output vector.

The attention is designed as 'h' times, considering it is also known as the multi-headed, so each repetition calculates as the head. The parameter 'W' is different each time q, k, also v undertake a linear layer transformation. Consequently, the outcome from the 'h' repetitions of scaled-dot attention transformation is linked, also the value acquired by additional linear layer transformation is the product of the MHA, by considering 'q' queries from matrix 'Q'. Furthermore, calculating the arrangement using general attention dot-product and to compute context vectors can be minimized as below equation.

$$Attention(Q,k,v) = softmax(Q * k^T) * v \tag{24}$$

The MHA model considers this as another step. Here k, v is plotted into vector spaces including the lower-dimensional and combining into 'K', 'V' matrices and by indicating weight matrices 'W', next the results are assigned to calculate the attention, note: the output of these representations indicating as 'head'.

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{25}$$

The heads are then combined and converted by a weight matrix as output (Vaswani et al., 2017), where 'mh' indicates multi-head.

$$mh(q,k,v) = concat(head_1, \ldots, head_h)W^O \tag{26}$$

Where parameter matrices are the projections. The outcomes of each head 'h' are also merged. Although, after over a linear projection level is computed, with the resulted output executing the SA once on the earlier (Q, K, V) matrices (Ambartsoumian & Popowich, 2018).

$$W_i^Q \in \mathbb{R}^{d_{model}*d_k}, W_i^K \in \mathbb{R}^{d_{model}*d_k}, W_i^V \in \mathbb{R}^{d_{model}*d_v}, W_i^O \in \mathbb{R}^{d_{model}*hd_v} \tag{27}$$

Various sets of q, k and v are obtained by the same original sequence inputs $S = (x_1, x_2, \ldots, x_n)$ embedding. From SA phases are executed on every set distinctly and the latest embeddings $v_i, v_j$ are developed for every set. Embeddings are then combined and dot-product with another learned matrix $Q$ is trained and jointly which moderates these combined embeddings into a single embedding for $Y_1, Y_2 \ldots Y_n$. This is generally denoted as MHA, where each 'v' indicating the head 'h' of the model SA.

**Question: "when is not falling asleep fast enough considered insomnia?"** The attention model splits the above sentence into multiple attention mechanisms. The MHA can have more than one attention mechanism. By considering the features extracted from the previous layer. In the above question related to the mental health issue, the attention will consider the vectors from the previous layer or embeddings layer, the given sentence will be stored in the vectorized format (a) and using the cosine similarity between the co-occurrence of words. The bold words related to multiple heads associated with the linear layer (b), MHA considers these heads randomly depending upon the relation between the words. Vectors with multiple heads that are scaled to the dot product attention (c), as it encompasses blocks of MHA, whereas the attention computing itself is scaled dot-product attention, these weights are passed to the hidden state (d) and combined all weights together and form $[W_{ij}]$ in contact layer (e) and passing to dense layer to get the hidden features from the previous layer, which resulted as that vector contains the combined information from the previous hidden layer. Essentially, considering various values from the previous linear layers before passing through contact and dense layers, so in order to combine the linear layers, constrain the values using the contact and dense layer.

While one attention head attends to the tokens that are relevant to each token, by multiple attention heads the model can learn to do this for different definitions of "relevance". Research has shown that many attention heads in transformers encode relevance relations that are interpretable, for example, there are attention heads for every token that attend mostly to the subsequent word or attention heads that mainly attend from verbs to their direct objects (Partaourides et al., 2020). Since transformer models have multiple attention heads, they have the possibility of capturing many levels and types of relevance relations, from surface-level to semantic level. The multiple outputs for the MHA layer are concatenated (f) to pass into the FNN layers. Although the single head attention appears to the tokens which are appropriate to every token, with MHA the mechanism can acquire to complete this for various explanations of "relevance". Considering the head-attentions for each token, appear frequently to the subsequent word and MHA executes multiple attention functions concurrently (XinSheng & Teng, 2019). The various outputs for the MHA layer are merged to pass into the FNN layers. Unlike single head attention, in MHA does not have to over-saturate on one single attention mechanism because it can spread the cognitive load on multiple ones. Additionally, as an alternative of executing SA, this paper suggests the MHA with dimension $d_{model}$ for (q, k,v) from the (Vaswani et al., 2017) suggested MHA, which executes attention 'h' times on projected combination of (q,k,v) of $d_{model}/h$ related to dimension matrices. In this, each layer consists of two sub-layers. Here, the first sub-layer denotes the multi-head SA method, whereas the next sub-layer is a general fully connected-FNN position-wise operation. To assign the remaining connections for every sub-layer, considered the normalization layer. The following diagram represents the neural-architecture of our proposed model Fig. 1.

Fig. 2 represents the structure of the MHA-BCNN model. Firstly, the data corpus is categorized into various documents. Wherein each word in the sentence represented as tokens, and these tokens are stored in a vectorized format, using the GloVe word-embeddings, the pre-trained values considered from Pennington et al. (2014), and the words are trained using the cosine similarity resulted as the correlation between the words, which indicated as pre-trained weightings and these including for our model training. As stored each word in a vectorized format and these vectors are passed to the BiLSTM layer, the advantage over LSTM is the bidirectional LSTM can pass the input features from both forward and backward direction, here at each stage storing the information at each RNN and further, passed the input values to MHA layer. In this layer, extracted the hidden features from the previous BiLSTM layer (Bahdanau et al., 2014) and form as a linear layer and resulting various linear layers corresponding to the hidden features, here each linear layer will have multiple heads and each head represents the position of the input features, thus the multiple heads (Eq. (26)). So, combined all these layers using the dense layer. Then, applying the convolution and global max pooling, for a filter size 'f' and further linked to the fully-connected softmax function in Algorithm 1, the output layer which results that (L,N,E), where 'L' indicates about target sequence length, where the batch size is 'N' embedding dimension is 'E' and resulted in the types of emotions concerning the inputs. The MHA permits the architecture to conjointly appear to information from various depiction subspaces at various points. The enhancement using the MHA having attention at each 'word', it can correlate with the other word which is far away from the particular word, unlike RNN need to compare with adjacent words rather than words which are far away from the particular word.

From the MHA-BCNN algorithm, where line 5 indicates calculating the cosine similarity between the word vectors. Where these trainings are fed into two independent LSTMs in line 8. Concatenating the hidden features from both forward and backward LSTMs in line 10. Similarly from line 12 feeding these features into the MHA layer and form as a linear layer, where the position of the input features represents as 'head'. In line 14 further linking to max-pooling contains filter size 'f' with softmax function which results to predict the class labels and for linking the predicted values with the suitable class and finally classification result is generated in the output layer.

**Algorithm 1**: MHA-BCNN Algorithm

1: **function** MentalHealth(questions)
2: $W \leftarrow$ StemAndTokenize(questions)
3: **for all** word $\in W$ **do**
4: Converting into equivalent word vectors
5: Through Glove V $\leftarrow [v_1, v_2, \ldots, v_n]$ with Eq. (1)
6: **end for**

(continued on next page)

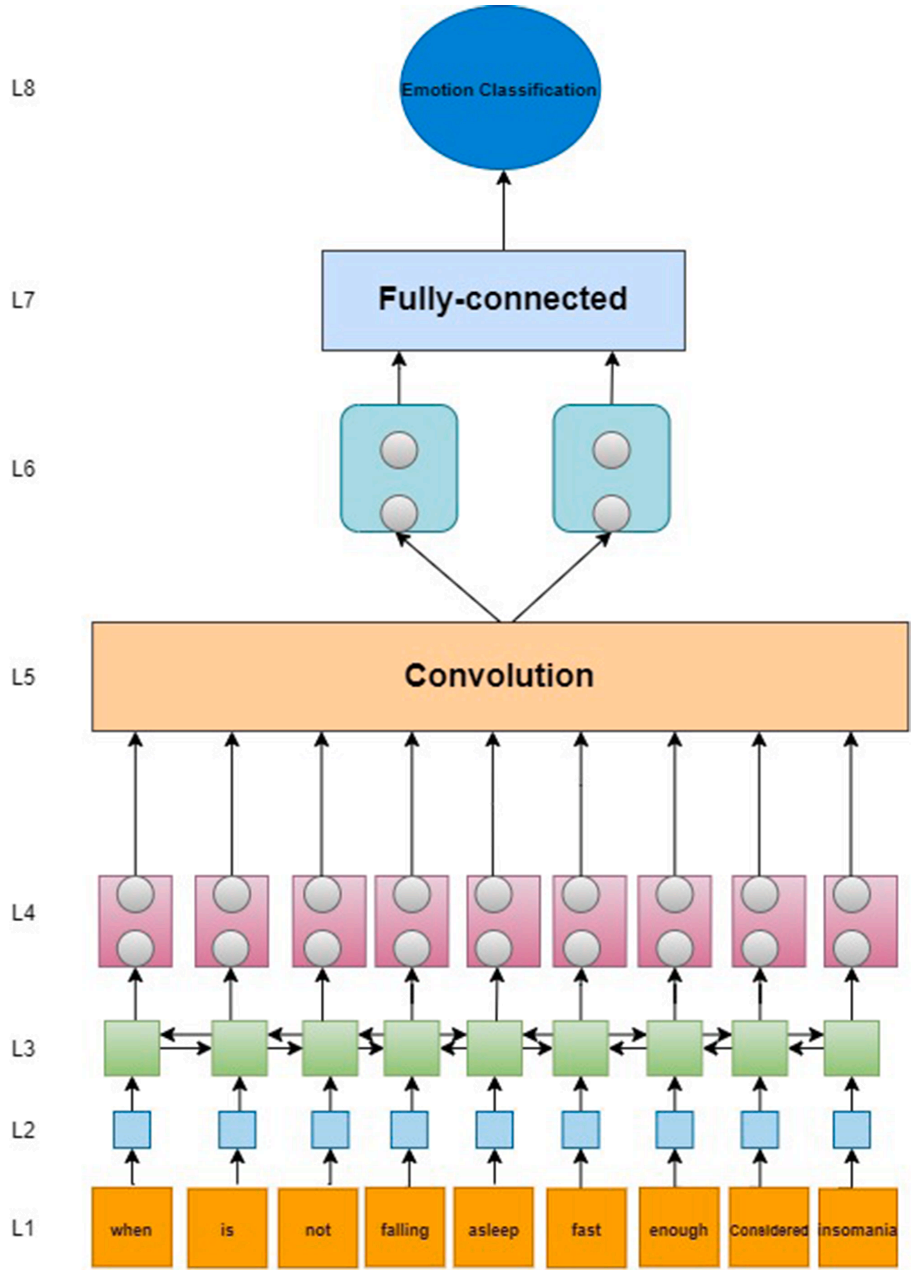



**Fig. 2.** The architecture of the proposed MHA-BCNN model. L1- word tokens; L2- word embeddings; L3- BiLSTM; L4- MHA layer; L5- convolution; L6- globalmaxpooling; L7- fully-connected layer; L8- emotion classification.

(*continued*)

**Algorithm 1**: MHA-BCNN Algorithm

```
        for each word vector ∈ V do
            Passing word vectors to the BiLSTM with Eqs. (9) and (10)
            It generates hidden features h_1, …, h_n with Eq. (16)
            h_t = (→h_{t−1}, ←h_{t−1}) Eq. (11)
            To spread the cognitive load by MHA
            mh(q, k, v) = concat(head_1, ., head_h) W^O
            Further linked to fully-connected softmax
             a_ij = softmax(x_j, i_t)
        end for
    end function
```

## 5. Experimental results

### 5.1. Dataset collection

Extracted the text related to mental health issues from both Webmd and Healthtap online medical healthcare platform which contains patient-doctor interaction, maintained by the medical professionals including the psychiatrists. The questions related to mental health were posted by the patients through email, then the clinicians assign the labels to each question. Our datasets Webmd about 2086 and Healthtap about 5328 questions respectively. Where applied the datasets on various DL models. We used the Google dump 6 billion tokens, 2014 pretrained GloVe embeddings (Pennington et al., 2014) to identify the most similar words about 100, 200, 300 dimensions. In addition to this, received the best performance using the GloVe embeddings 300 dimensions to train our model. We used GloVe embedding to capture the long sequence of data than word2vec preferable only for the short texts. The overall two datasets are distributed into various phases like training as 75%, 25% for testing and validation purpose. Both the Binary class entropy and Adam is the best classification and optimizer to identify the emotions for long-text documents.

#### 5.1.1. Discretion of the dataset

From the (Table 2) underlined words represent the multi-emotion labels in the question associated with word emotion.

#### 5.1.2. Data-preprocessing

We extracted the text considering the mental-health issues related questions from both Healthtap and Webmd online healthcare platform. We removed the stop-words, punctuations, special characters, hashtags and converted them to lowercase letters, in order to get the appropriate tokens applied to the preprocessing procedure. Considered the max length sequence words as 500 and the length less than our consideration can be padded using the zero-padding procedure, here including an additional zero to the sequence. We sent the tokens to the word embedding, where we used the GloVe vectors. The following (Figs. 3–6) show the pool of tokens related to mental health care from both datasets.

### 5.2. Results and discussions

To extract the emotions from long-text documents, we built and compared classifiers like CNN, LSTM, LSTM-CNN, BiLSTM, BiLSTM-CNN, GRU-CNN, GRU-LSTM, GRU-LSTM-CNN, GRU-BiLSTM, GRU-BiLSTM-CNN, MHA-CNN, MHA-LSTM, MHA-LSTM-CNN, MHA-BiLSTM, including other state-of-art methods and our model MHA-BCNN with GloVe embeddings (100, 200, 300 dimensions). To check the overall accuracy of the system was calculated to various iterations like 6, 8, 10, and batch size about 100, as our model about multi-labeled classification with Adam optimizer and achieved accuracy about 97.8% than previous results.

#### 5.2.1. Performance measures

The foremost intention for the MHA-BCNN to have further attention between the words than CNN, RNN, LSTM, BiLSTM, GRU models where the MHA-BCNN mechanism can also work better for long-text documents. By taking the number of attention heads as 8, the hidden layer size in the feed-forward network inside the transformer is 32. BiLSTM dimensions with 128, 256 and global max-pooling with filter sizes 2, 3 respectively and Rectified Linear Unit (ReLU) for CNN model and at last combine with a softmax activation function. Our model resulted in about 97.8 % accuracy using the Healthtap MHA-BCNN model which outperforms all other DL models. Our model works better for both training and validation purposes unlike all other models and also having better precision and recall, achieving better accuracy and F1-score metrics than any other previous research works, considering long-sequence data as an input. For this MHA-BCNN model, we considered 10 epochs. Where in each iteration step the model improves its performance by considering more trainable parameters. At each epoch, the model performs better, so the model learns itself at each step and improves its accuracy (Table 3). The following (Figs. 7–10) show the performance of our proposed model for both Webmd and Healthtap datasets.

**Table 2**
The mental health related questions posted by the patients.

| Question | Emotion Label |
|---|---|
| "Zoloft (sertraline) has helped my anxiety but now I feel in between emotions. I don't get overly excited like I use to and I don't get down as much. is this normal?" | **Anxiety** |
| "When is not falling asleep fast enough considered insomnia?" | **Insomania** |
| "Zoloft (sertraline) helped with normal ocd but caused intrusive unwanted thoughts after a month. what can I do now?" | **OCD** |

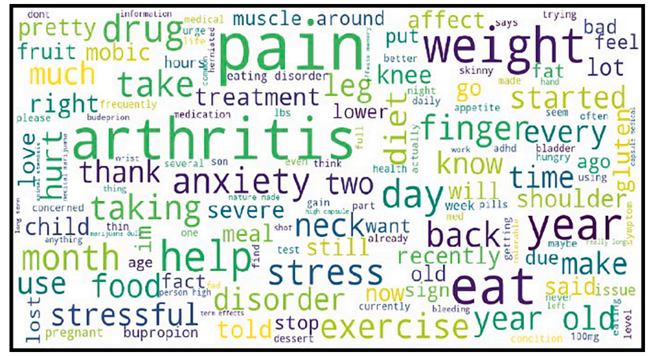



**Fig. 3.** Words cloud related to mental-health texts (Webmd).

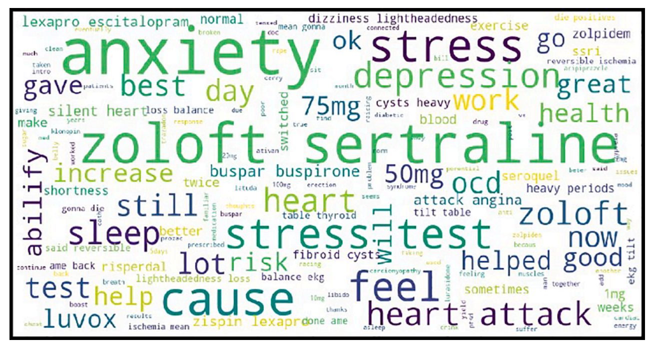



**Fig. 4.** Words cloud related to mental-health texts (Healthtap).

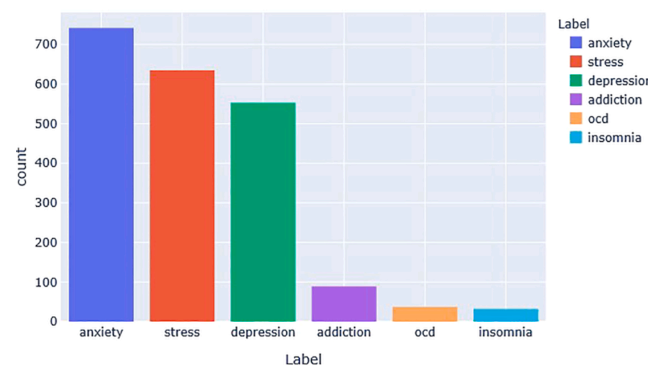



**Fig. 5.** Emotion-wise amount of words (Webmd).

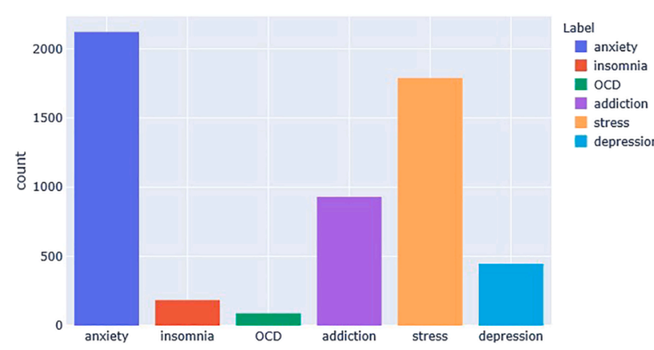



**Fig. 6.** Emotion-wise amount of words (Healthtap).

**Table 3**
The below experimental results show the overall performance metrics of DL models with our proposed model below.

| DL Model | GloVe Dimensions | Webmd | Health-tap | Performance |
|---|---|---|---|---|
| CNN | 100 | 0.765 | 0.906 | F |
| | | 0.926 | 0.969 | A |
| | | 0.815 | 0.921 | P |
| | | 0.722 | 0.892 | R |
| | 200 | 0.799 | 0.906 | F |
| | | 0.936 | 0.969 | A |
| | | 0.837 | 0.916 | P |
| | | 0.766 | 0.896 | R |
| | 300 | 0.799 | 0.914 | F |
| | | 0.934 | 0.971 | A |
| | | 0.818 | 0.920 | P |
| | | 0.782 | 0.908 | R |
| LSTM | 100 | 0.778 | 0.891 | F |
| | | 0.931 | 0.963 | A |
| | | 0.835 | 0.887 | P |
| | | 0.730 | 0.874 | R |
| | 200 | 0.795 | 0.892 | F |
| | | 0.934 | 0.965 | A |
| | | 0.824 | 0.891 | P |
| | | 0.768 | 0.889 | R |
| | 300 | 0.775 | 0.895 | F |
| | | 0.929 | 0.966 | A |
| | | 0.821 | 0.899 | P |
| | | 0.734 | 0.892 | R |
| LSTM-CNN | 100 | 0.791 | 0.914 | F |
| | | 0.933 | 0.971 | A |
| | | 0.822 | 0.932 | P |
| | | 0.762 | 0.908 | R |
| | 200 | 0.808 | 0.921 | F |
| | | 0.938 | 0.974 | A |
| | | 0.832 | 0.933 | P |
| | | 0.786 | 0.912 | R |
| | 300 | 0.814 | 0.924 | F |
| | | 0.940 | 0.975 | A |
| | | 0.845 | 0.935 | P |
| | | 0.786 | 0.918 | R |
| BiLSTM | 100 | 0.779 | 0.908 | F |
| | | 0.928 | 0.969 | A |
| | | 0.806 | 0.914 | P |
| | | 0.754 | 0.902 | R |
| | 200 | 0.803 | 0.909 | F |
| | | 0.936 | 0.970 | A |
| | | 0.827 | 0.913 | P |
| | | 0.780 | 0.906 | R |
| | 300 | 0.806 | 0.918 | F |
| | | 0.936 | 0.973 | A |
| | | 0.817 | 0.928 | P |
| | | 0.796 | 0.908 | R |
| BiLSTM-CNN | 100 | 0.804 | 0.892 | F |
| | | 0.935 | 0.961 | A |
| | | 0.844 | 0.895 | P |
| | | 0.768 | 0.880 | R |
| | 200 | 0.805 | 0.891 | F |
| | | 0.936 | 0.963 | A |
| | | 0.816 | 0.903 | P |
| | | 0.794 | 0.892 | R |
| | 300 | 0.816 | 0.899 | F |
| | | 0.940 | 0.964 | A |
| | | 0.842 | 0.906 | P |
| | | 0.792 | 0.888 | R |
| GRU-CNN | 100 | 0.760 | 0.892 | F |
| | | 0.922 | 0.965 | A |
| | | 0.782 | 0.915 | P |
| | | 0.740 | 0.870 | R |
| | 200 | 0.781 | 0.898 | F |
| | | 0.930 | 0.966 | A |
| | | 0.819 | 0.904 | P |
| | | 0.746 | 0.892 | R |
| | 300 | 0.793 | 0.911 | F |
| | | 0.933 | 0.971 | A |

**Table 3** (*continued*)

| DL Model | GloVe Dimensions | Webmd | Health-tap | Performance |
|---|---|---|---|---|
| | | 0.816 | 0.922 | P |
| | | 0.772 | 0.902 | R |
| GRU-LSTM | 100 | 0.752 | 0.913 | F |
| | | 0.920 | 0.970 | A |
| | | 0.780 | 0.915 | P |
| | | 0.726 | 0.905 | R |
| | 200 | 0.764 | 0.915 | F |
| | | 0.927 | 0.971 | A |
| | | 0.826 | 0.919 | P |
| | | 0.712 | 0.912 | R |
| | 300 | 0.787 | 0.919 | F |
| | | 0.932 | 0.973 | A |
| | | 0.823 | 0.926 | P |
| | | 0.754 | 0.913 | R |
| GRU-LSTM-CNN | 100 | 0.774 | 0.898 | F |
| | | 0.928 | 0.966 | A |
| | | 0.811 | 0.904 | P |
| | | 0.742 | 0.892 | R |
| | 200 | 0.810 | 0.902 | F |
| | | 0.939 | 0.968 | A |
| | | 0.835 | 0.913 | P |
| | | 0.781 | 0.893 | R |
| | 300 | 0.809 | 0.908 | F |
| | | 0.937 | 0.970 | A |
| | | 0.825 | 0.921 | P |
| | | 0.794 | 0.896 | R |
| GRU-BiLSTM | 100 | 0.757 | 0.908 | F |
| | | 0.923 | 0.969 | A |
| | | 0.800 | 0.909 | P |
| | | 0.720 | 0.905 | R |
| | 200 | 0.793 | 0.916 | F |
| | | 0.932 | 0.972 | A |
| | | 0.810 | 0.919 | P |
| | | 0.764 | 0.913 | R |
| | 300 | 0.800 | 0.921 | F |
| | | 0.936 | 0.974 | A |
| | | 0.841 | 0.932 | P |
| | | 0.778 | 0.910 | R |
| GRU-BiLSTM-CNN | 100 | 0.787 | 0.912 | F |
| | | 0.930 | 0.969 | A |
| | | 0.805 | 0.920 | P |
| | | 0.770 | 0.894 | R |
| | 200 | 0.807 | 0.918 | F |
| | | 0.937 | 0.971 | A |
| | | 0.832 | 0.927 | P |
| | | 0.784 | 0.914 | R |
| | 300 | 0.794 | 0.920 | F |
| | | 0.933 | 0.973 | A |
| | | 0.817 | 0.931 | P |
| | | 0.772 | 0.916 | R |
| MHA-CNN | 100 | 0.778 | 0.903 | F |
| | | 0.929 | 0.968 | A |
| | | 0.811 | 0.905 | P |
| | | 0.748 | 0.902 | R |
| | 200 | 0.791 | 0.914 | F |
| | | 0.933 | 0.971 | A |
| | | 0.824 | 0.917 | P |
| | | 0.760 | 0.912 | R |
| | 300 | 0.789 | 0.926 | F |
| | | 0.932 | 0.974 | A |
| | | 0.822 | 0.929 | P |
| | | 0.760 | 0.924 | R |
| MHA-LSTM | 100 | 0.793 | 0.908 | F |
| | | 0.935 | 0.970 | A |
| | | 0.840 | 0.921 | P |
| | | 0.752 | 0.894 | R |
| | 200 | 0.796 | 0.911 | F |
| | | 0.933 | 0.971 | A |
| | | 0.819 | 0.924 | P |
| | | 0.774 | 0.904 | R |
| | 300 | 0.794 | 0.927 | F |
| | | 0.934 | 0.975 | A |

(*continued on next page*)

**Table 3** (*continued*)

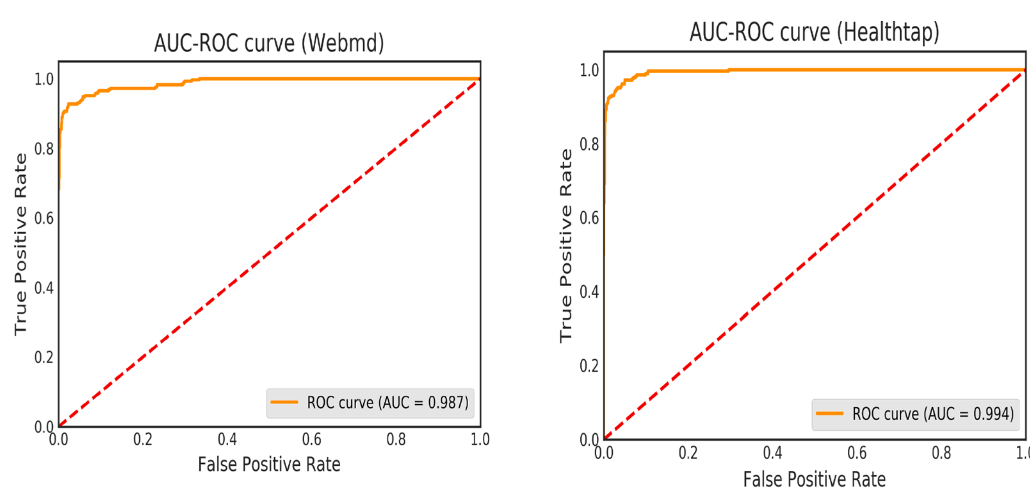
| DL Model | GloVe Dimensions | Webmd | Health-tap | Performance |
|---|---|---|---|---|
| | | 0.825 | 0.930 | P |
| | | 0.766 | 0.926 | R |
| MHA-LSTM-CNN | 100 | 0.800 | 0.919 | F |
| | | 0.937 | 0.972 | A |
| | | 0.839 | 0.924 | P |
| | | 0.750 | 0.910 | R |
| | 200 | 0.814 | 0.922 | F |
| | | 0.942 | 0.974 | A |
| | | 0.852 | 0.925 | P |
| | | 0.778 | 0.920 | R |
| | 300 | 0.813 | 0.930 | F |
| | | 0.941 | 0.976 | A |
| | | 0.863 | 0.935 | P |
| | | 0.768 | 0.926 | R |
| MHA-BiLSTM | 100 | 0.796 | 0.910 | F |
| | | 0.934 | 0.970 | A |
| | | 0.823 | 0.916 | P |
| | | 0.771 | 0.904 | R |
| | 200 | 0.804 | 0.915 | F |
| | | 0.937 | 0.972 | A |
| | | 0.841 | 0.926 | P |
| | | 0.772 | 0.906 | R |
| | 300 | 0.818 | 0.919 | F |
| | | 0.940 | 0.973 | A |
| | | 0.836 | 0.925 | P |
| | | 0.802 | 0.910 | R |
| MHA-BCNN | 100 | 0.822 | 0.936 | F |
| | | 0.939 | 0.973 | A |
| | | 0.859 | 0.943 | P |
| | | 0.794 | 0.930 | R |
| | 200 | 0.840 | 0.945 | F |
| | | 0.943 | 0.976 | A |
| | | 0.864 | 0.946 | P |
| | | 0.832 | 0.936 | R |
| | **300** | **0.921** | **0.953** | F |
| | | **0.962** | **0.978** | A |
| | | **0.946** | **0.967** | P |
| | | **0.916** | **0.942** | R |

**F**: F1-score.
**A**: Accuracy.
**P**: Precision.
**R**: Recall.

### 5.2.2. *Evaluation and Comparison*

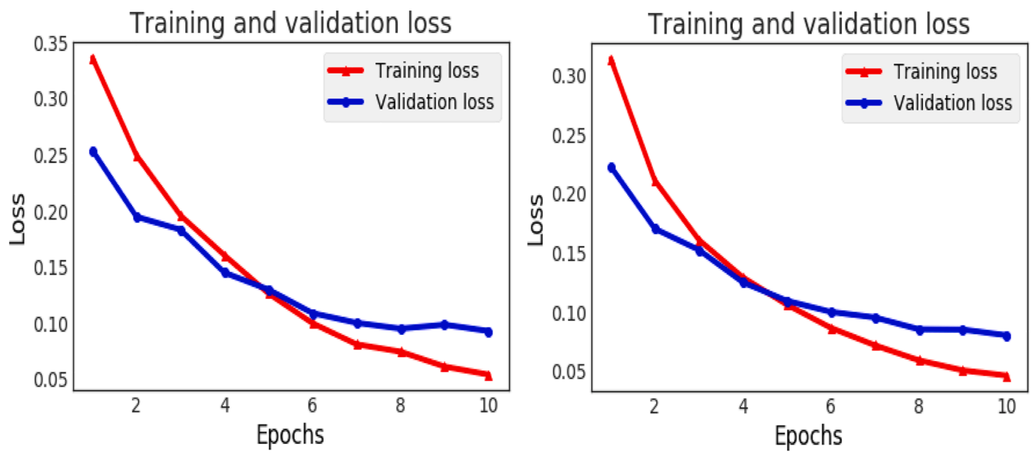
We compare our model with other state-of-art methods as mentioned below. Our model shows an outstanding performance than the erstwhile methods.

- **CMA-MemNet:** To maintain the semantic relations at the word level. Proposed CMA-MemNet (Zhang et al., 2020) is a self-attention-based memory network to capture the context information with semantic relations between the words. Considered the 6 convolution processes and dropout is 0.2, batch size is 20, number of epochs as 40, learning rate about $7 \times 10^{-5}$, the $L_2$ regularization as $1 \times 10^{-5}$, number of convolution filters is 300.
- **SENN:** To capture semantic information with emotion recognition by the SENN model (Batbaatar et al., 2019). This model is a combination of two sub-networks like BiLSTM for context information and other-side CNN with emotional features. With learning rate about 0.001, batch size is 128, hidden dimension for CNN is 256, number of filters is 100, filter sizes considered as 3, 4 and 5, dropout is 0.5 for using SENN model.
- **Single-layered BiLSTM:** For detecting text-based emotions, proposed single-layered BiLSTM (Hameed & Garcia-Zapirain, 2020) with embedding features. By using the optimum settings like learning rate as 0.0001, drop out is 0.3, batch size is 64, BiLSTM nodes as 16, maximum length as 40, epochs as 70, optimizer as RMSProp, loss function with cross-entropy and implemented on $L_2$ regularization.
- **MAN:** Proposed MAN (Zhang & Lu, 2019) for text representation in aspect terms to obtain sequence information. To capture the hidden features at various positions of words at the aspect level. With the MAN model, they included the settings like dropout rate as 0.5, also used the regularization parameter $e$ with 0.1, where the learning rate considered as $5^{-5}$ and $10^{-5}$ for $L_2$ regularization item.
- **Attention-Base C-BiLSTM:** Proposed Attention-Base C-BiLSTM (Ahmad, Asghar, Alotaibi, & Khan, 2020) to classify the texts with various emotional states. For the model C-BiLSTM, they extracted the max-features as 1000, batch size is 32, kernel sizes consists of $2 \times 2$ and $3 \times 3$, the maximum length is 53. Also, considered the dropout bit size as 0.5, BiLSTM unit size between 30 to 80 for six attention based C-BiLSTM mechanisms, 8 number of hidden layers, attention layer value is 1, number of epochs as 8, they implemented on softmax activation function, number of filters is considered as 16, 32 and 64, convolution layer number with value 1 and embedding dimension is 128.
- **CNN_Text_Word2vec:** To learn the text features with emotion detection from CNN_Text_Word2vec model (Xu et al., 2020). They



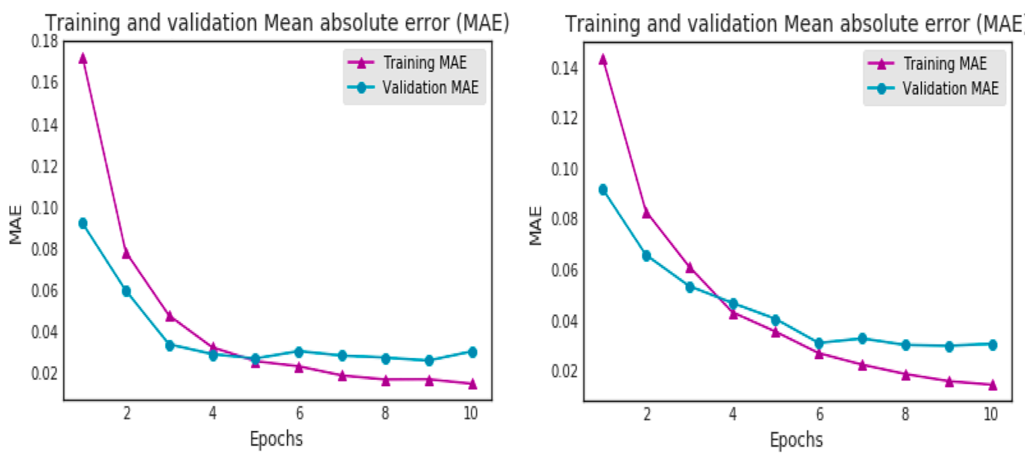

(a) (b)

**Fig. 7.** AUC-ROC curve (a) Webmd (b) Healthtap.



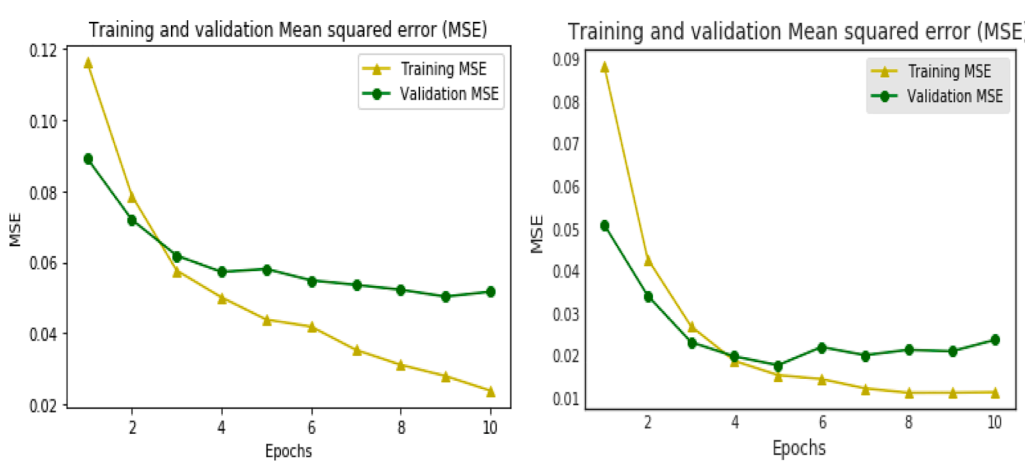

(a) (b)

**Fig. 8.** Loss function (a) Webmd (b) Healthtap.




(a) (b)

**Fig. 9.** Mean absolute error (a) Webmd (b) Healthtap.




(a) (b)

**Fig. 10.** Mean squared error (a) Webmd (b) Healthtap.

used the settings like 1-max pooling size with pooling operator, learning rate as 0.001, filter sizes like 3, 4 and 5, drop out value is about 0.5, the batch size is 50, they considered the word vector dimension during the training process as 200.

- **COMV:** Proposed model COMV (Xiao, Wei, Mao, & Wang, 2019) for emotion-oriented extraction features, which is extracted as independent and potential emotion-clauses. With dropout rate as 0.8, batch size is 32, size of word embeddings and the hidden units as 30 at a learning rate of $10^{-3}$.
- **EvoMSA:** Proposed EvoMSA model (Estrada et al., 2020) for classification of the opinions which reflects the emotional state. Mentioned two corpus sentiText and eduSERE. This is the evolutionary approach with a comparison of ML and DL classifiers. Conducted experiments on q-grams of characters with sizes 2, 3, 4, 5, and 6.
- **VC(LR-SGD):** Proposed LR-SGD model for voting classifier and analyzed the public opinions. The model VC(LR-SGD) (Yousaf et al., 2021) with TF-IDF features for detecting the emotion recognition and also implemented on seven ML models. With SGD optimizer, for ProbSGD-Pos and ProbSGD-Neg with probability scores 0.997 and 0.002 respectively.
- **VCPCNN:** To maintain the semantic features from word vectors, proposed VCPCNN (Dong et al., 2020) with pooling convolutional network to capture the local features and used the word embeddings dimensions. Conducted the experiments on 1D wide convolution, 2D wide convolution with filter sizes $3 \times 1$ and $5 \times 1$.
- **CNN with GA:** Proposed a model with convolution with genetic algorithm(GA) (Ishaq, Asghar, & Gillani, 2020) for extracting the emotions from the texts. Semantic features are extracted by tuning the CNN parameter with multi-objective function like GA, also used the word embeddings. By using the CNN model they considered the convolution with 3 region sizes and 2 filters for each region size with a total of 6 filters. Also, used the 1-max pooling with softmax function on 6 univariate vectors and concatenated as a single feature vector.
- **CNN-BiLSTM:** Proposed CNN-BiLSTM (Rhanoui et al., 2019) with Doc2vec for extracting the opinions from long texts at document-level analysis. By using the CNN-BiLSTM model implemented on four layers of convolution with filter sizes 2, 3, 4 and 5, considered the ReLU function, the value of the batch size is 32 and 64, they used SGD optimizer, also included the Adam optimizer and compared with 6, 8 and 10 epochs.
- **CE-B-MHA:** Proposed a model for analyzing the emotional tendency with text classification. By CE-B-MHA model (Lin et al., 2020) aims to perform a comparison between feature vectors and labeled vectors. Considered various sample sizes and they received the better performance with a sample size greater than 35.
- **TextCNN:** By using TextCNN model with TF-IDF features (Song et al., 2019) for text-based classification. By using TextCNN model compared with various filter sizes like 2, 3 and 4, used the pooling layer to compress the features, they included the softmax activation function. Also, KNN and TextCNN were used for text classification.

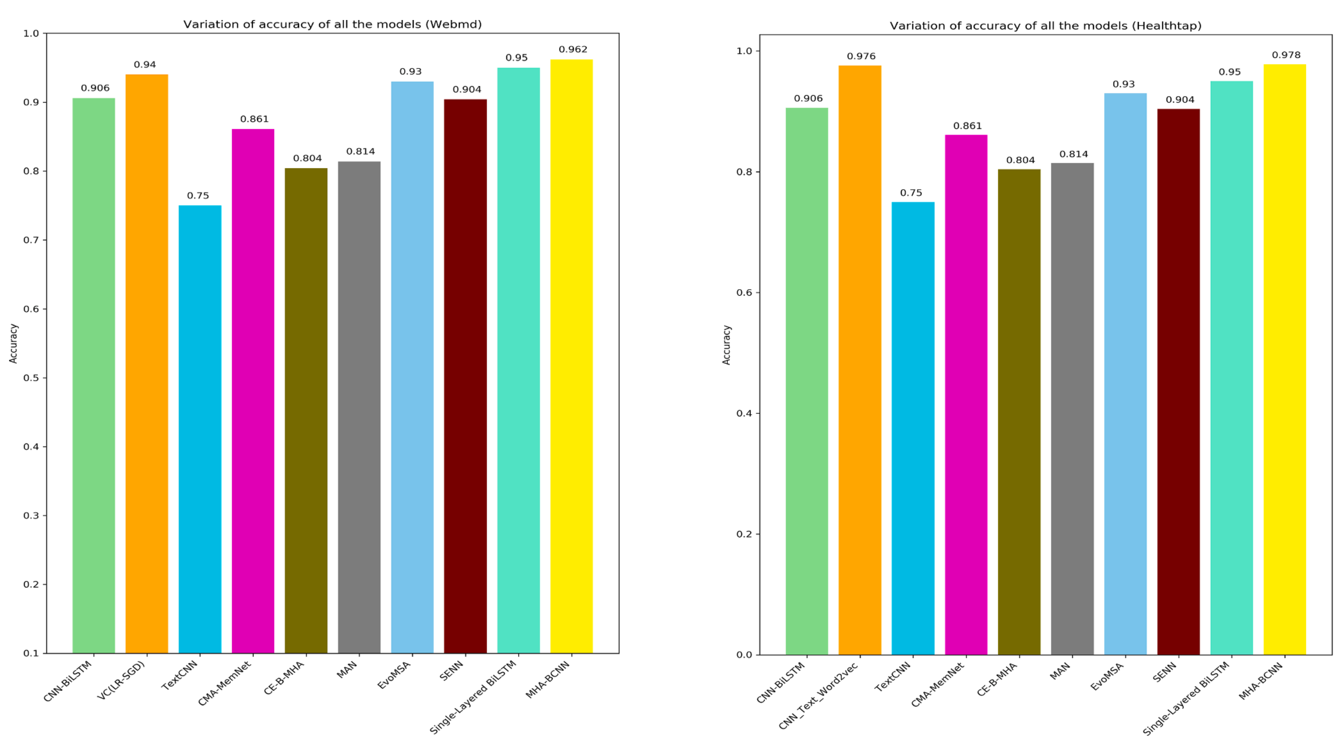
The (Figs. 11–14) show the performance metrics like accuracy, F1-score, precision, and recall with comparison to other state-of-art methods. Our MHA-BCNN model gave the best results than the other models (for detailed results Table 3).

In (Table 3) the MHA-BCNN model using GloVe embeddings dimension 300 for both Webmd and Healthtap shows the best performance than the other DL models.

**Macro F1-score** The following represents the top 5 best performance models for macro F1-score, it calculates the metrics for each DL model and their unweighted mean (Sokolova & Lapalme, 2009). This includes the label disparity into consideration.

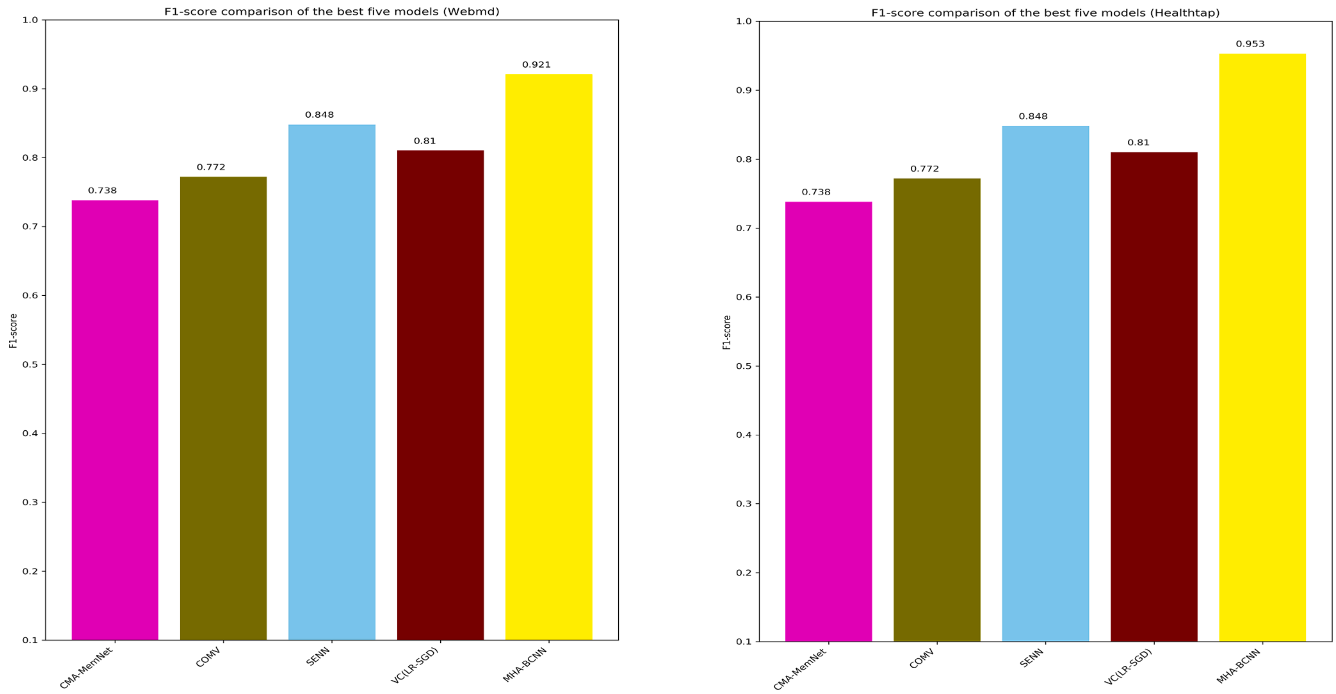
$$Macro_{F1\text{-}score} = [2 * (Macro\text{-}Precision * Macro\text{-}Recall)/(Macro\text{-}Precision + Macro\text{-}Recall)] \quad (28)$$

In (Table 4) model MHA-BCNN with GloVe embeddings dimension 300 having about 0.783% macro F1-score for Webmd and Healthtap about 0.895 % score better than the previous DL models.



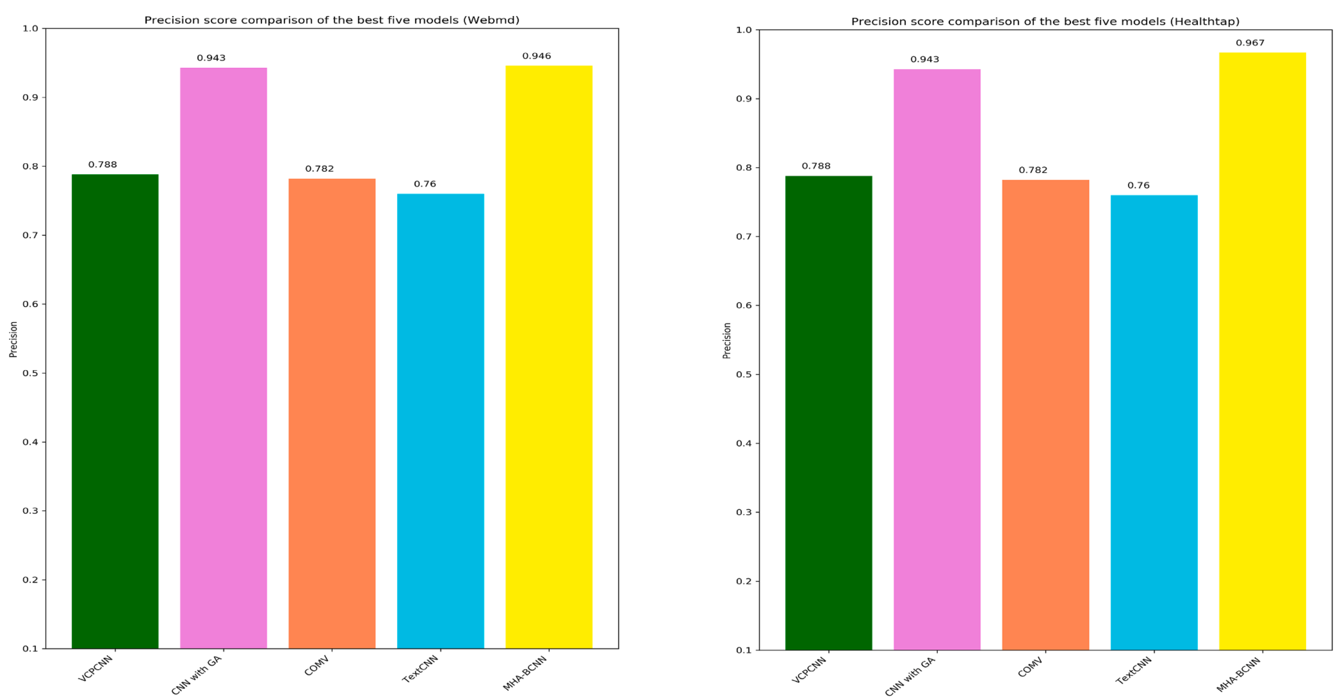

(a) (b)

**Fig. 11.** Variation of the accuracy of all the models (a) Webmd (b) Healthtap.




(a) (b)

**Fig. 12.** Variation of the F1-score from best five models (a) Webmd (b) Healthtap.

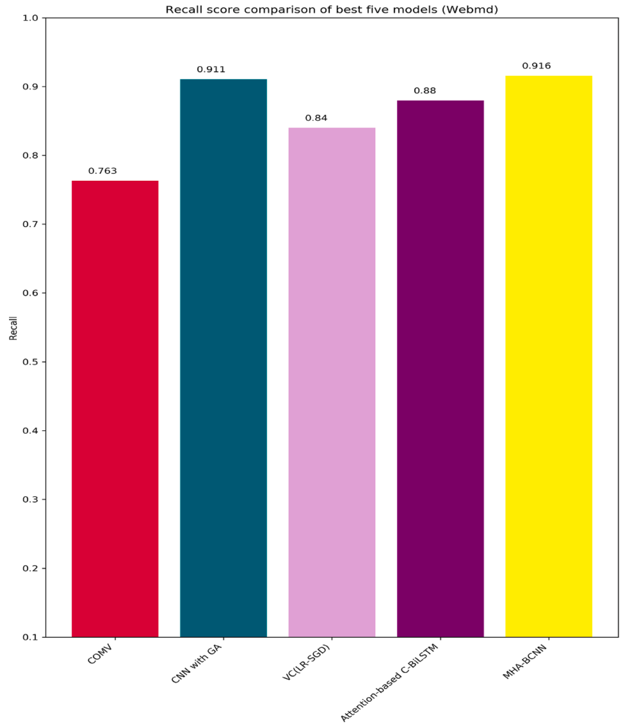



(a) (b)

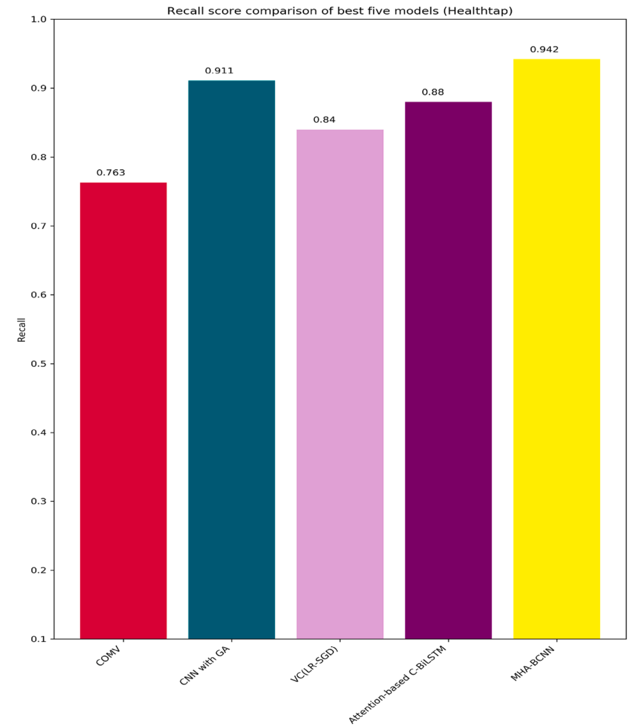
**Fig. 13.** Variation of the precision score from best five models (a) Webmd (b) Healthtap.

### *5.3. Optimizer*

Our model implemented on Adam optimizer that can be observed by a grouping of momentum with RMSprop and stochastic gradient descent (SGD), similar to RMSprop it considers the squared gradients to measure the learning rate (Li et al., 2019). Also, Adam takes the benefit of momentum included the gradient moving average in place of the gradient for instance SGD with momentum (Kingma & Ba, 2015). By computing the loss for each class independently and computed as:




(a)




(b)

**Fig. 14.** Variation of the recall score from best five models (a) Webmd (b) Healthtap.

**Table 4**
The following table shows the experimental result of macro F1-score.

| Model | Webmd | Health-tap |
|---|---|---|
| LSTM-CNN | 0.720 | 0.810 |
| BiLSTM | 0.760 | 0.764 |
| BiLSTM-CNN | 0.773 | 0.840 |
| GRU-BiLSTM-CNN | 0.747 | 0.843 |
| MHA-BCNN | **0.783** | **0.895** |

$$X = -\Sigma_{i=1}^{k} P(X)\log Q(X) \tag{29}$$

Using the dot product between the probability of class P(X) and target with a probability of class Q(X) in prediction. Let 'm' be the moment, '$X$' be the random variable it can be denoted as $m_k = E[X^k]$, where $k^{th}$ moment related to a random variable is illustrated as 'E' for the expected value of a variable to the power of 'k'. By calculating the values from the past squared gradients, $m_k, v_k$ computes the mean and variance of the gradients 'g' at $k^{th}$ moment correspondingly.

$$m_k = \beta_1 m_k - 1 + (1-\beta_1)g_k \tag{30}$$

$$m_k = m_K/1-\beta_1^K \tag{31}$$

$$v_k = \beta_1\ m_k - 1 + (1-\beta_2)g_2^k \tag{32}$$

$$v_k = v_k/1-\beta_2^k \tag{33}$$

From the Adam optimizer with default values $\beta_1 = 0.9, \beta_2 = 0.999$ and $10^{-8}$ for $\in$, where the default values used to update the parameters likewise RMSprop and Adadelta (Ren et al., 2020). Which resulted as best optimizer for dealing with multi-labeled classification especially with long text documents.

#### 5.3.1. Dropouts

Removing the data overfitting is a stimulating task using the DL, so for each sub-layer, implemented the dropouts. The large weights in a neural network (NN) are a mark of a further intricate network which includes the training data overfitting (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The meek and active regularization scheme is the network with probabilistically reducing the out nodes. A huge network with extra training and considering the need for constraint weight is recommended when choosing the dropout. The Dropout is the method, where randomly opted neurons are unheeded through training. These are randomly dropped out, which indicates that their involvement in the initiation of temporally ensuing neurons was detached to the forward pass and any updates of weights are not implemented on the backward pass of a neuron. Also, having the dropout to the quantities of the word embeddings and the encodings positions in stacks of both encoder and decoder. Where every unit is retained with a stable probability and the learning rate is calculated as $P_{drop} = 0.1$.

### 5.4. Evaluation metrics

There are different metrics used to compute the performance relations, that can be examined and founded on the parameters likewise accuracy, F1-score, macro-F1, precision, and recall. The performance assessment plays an important part in accuracy estimation through an emotion analysis.

$$Acc = CP/TP \tag{34}$$

Where Acc is accuracy, CP stands for correct predictions and TP stands for total predictions.

**Precision**

$$P = TruePos/(TruePos + FalsePos) \tag{35}$$

**Recall**

$$R = TruePos/(TruePos + FalseNeg) \quad (36)$$

**F1-score**

The F1-score is the furthermost significant measure and considers how our data is executing. F-measure is a widespread F1-score and it measures the weighted harmonic mean of precision(P) and recall(R).

$$F_1 score = [2 * (P * R)/(P + R)] \quad (37)$$

## 6. Conclusion

In this paper, we addressed the problems related to large sequential text information with emotion analysis using psychiatrist text. We noticed these problems are the most significant aspects when dealing with emotion recognition using long text data on mental health-care-oriented texts. These problems result that a new direction for the exploration of emotion analysis on mental health care using the DL models. From our literature survey, we understood that emotions are different from different situations in an online mental-health-oriented text. So, need to capture that sequential text using MHA for better results. In this, we extracted the emotions from psychiatric data with long text sequences and reduced loss function using the DL models. In this research, we proposed the MHA-BCNN model for emotion analysis with GloVe Embedding. Our model gives decent outcomes to the long sequence text, meanwhile, it advantages from the MHA-BCNN capability to mining the features also to solve the bidirectional long-term dependencies from the text. Moreover, GloVe embeddings procedures the illustration of the text at the document level, which is further appropriate for the long-text document classification, distinct to the conventional short text organizing the emblematic of the online healthcare platform. The model is trained on a dataset from mental-health questions posted by the patients from Webmd and Healthtap websites. For both training and validation, by associating with various GloVe Wikipedia-2014 embeddings, dimensions as 100, 200, and 300 and DL models like CNN, LSTM, CNN-LSTM, BiLSTM, BiLSTM-CNN, GRU-CNN, GRU-LSTM, GRU-LSTM-CNN, GRU-BiLSTM, GRU-BiLSTM-CNN, MHA-CNN, MHA-LSTM, MHA-LSTM-CNN, MHA-BiLSTM with proposed model MHA-BCNN with various models GloVe embeddings and we concluded by experimental results that our model MHA-BCNN by GloVe-300 with Healthtap dataset is achieved the finest accuracy about 97.8% than any other DL models including state-of-the-art methods.

As for future work direction, we will focus on extracting the text-based emotions with temporal changes. In addition, we will recognize the emotion-wise mental-health posts on social media. Also, study on lexicon-based feature extraction like POS tags and another direction like working on the topic modeling concept. Further, we will extend our investigation by considering more transformer-based techniques.

## CRediT authorship contribution statement

**Kodati Dheeraj:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization, Project administration. **Tene Ramakrishnudu:** Supervision, Validation, Project administration, Writing - review & editing.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

Ahmad, S., Asghar, M. Z., Alotaibi, F. M., & Khan, S. (2020). Classification of poetry text into the emotional states using deep learning technique. *IEEE Access. 8*, 73865–73878.

Al-Abri, R., & Al-Balushi, A. (2014). Patient satisfaction survey as a tool towards quality improvement. *Oman Medical Journal. 29*(1), 3–7.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*.

Almeida, A. M., Cerri, R., Paraiso, E. C., Mantovani, R. G., & Junior, S. B. (2018). Applying multi-label techniques in emotion identification of short texts. *Neurocomputing, 320*, 35–46.

Ambartsoumian, A.& Popowich, F. (2018). Self-attention: A better building block for sentiment analysis neural network classifiers. In (pp. 130-139).

Bahdanau, D., Cho, K.& Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. ArXiv, 1409.

Bahgat, M., Wilson, S.& Magdy, W. (2020). Towards using word embedding vector space for better cohort analysis. Proceedings of the International AAAI Conference on Web and Social Media, 14(1), 919–923.

Bai, Y., Lin, C.-C., Chen, J.-Y., & LIU, W.-C. (2001). Virtual psychiatric clinics. *The American journal of psychiatry, 158*, 1160–1161.

Balahur, A., Hermida, J., & Montoyo, A. (2012). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems, 53*, 742–753.

Barron Estrada, M., Zatarain Cabada, R., Oramas, R., & Graff, M. (2020). Opinion mining and emotion recognition applied to learning environments. *Expert Systems with Applications, 150*, Article 113265.

Batbaatar, E., Li, M., & Ryu, K. H. (2019). Semantic-emotion neural network for emotion recognition from text. *IEEE Access, 7*, 111866–111878.

Boiy, E., & Moens, M.-F. (2009). A machine learning approach to sentiment analysis in multilingual web texts. *Information Retrieval Journal, 12*, 526–558.

Brailovskaia, J., & Margraf, J. (2016). Comparing facebook users and facebook non-users: Relationship between personality traits and mental health variables - an exploratory study. *PLOS ONE, 11*, Article e0166999.

Brueckner, R., & Schulter, B. (2014). Social signal classification using deep blstm recurrent neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 4823–4827).

Chatterjee, A., Gupta, U., Chinnakotla, M., Srikanth, R., Galley, M., & Agrawal, P. (2018). Understanding emotions in text using deep learning and big data. *Computers in Human Behavior, 93*.

Chen, G., Ye, D., Xing, Z., Chen, J., & Cambria, E. (2017). Ensemble application of convolutional and recurrent neural networks for multi-label text categorization. In *2017 international joint conference on neural networks (ijcnn)* (pp. 2377–2383).

Chen, S., Guo, X., Wu, T., & Ju, X. (2020). Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. *Information Processing & Management, 57*(5), Article 102253.

Chen, Y., Lv, Y., Wang, X., Li, L., & Wang, F. (2019). Detecting traffic information from social media texts with deep learning approaches. *IEEE Transactions on Intelligent Transportation Systems, 20*(8), 3049–3058.

Choi, H., & Lee, Y. (2020). Deep learning based response generation using emotion feature extraction. In *2020 IEEE international conference on big data and smart computing (bigcomp)* (pp. 255–262).

Cox, C., Moscardini, E., Cohen, A., & Tucker, R. (2020). Machine learning for suicidology: A practical review of exploratory and hypothesis-driven approaches. *Clinical Psychology Review, 82*, Article 101940.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* In Naacl-hlt.

Dong, M., Li, Y., Tang, X., Xu, J., Bi, S., & Cai, Y. (2020). Variable convolution and pooling convolutional neural network for text sentiment classification. *IEEE Access, 8*, 16174–16186.

Fei, Z., Yang, E., Li, D., Butler, S., Ijomah, W., Li, X., & Zhou, H. (2020). Deep convolution network based emotion analysis towards mental health care. *Neurocomputing, 388*, 212–227.

Fesseha, A., Xiong, S., Emiru, E. D., Diallo, M., & Dahou, A. (2021). Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya. *Information, 12*, 52.

Gao, K., Xu, H., Gao, C., Hao, H., Deng, J., & Sun, X. (2018). Attention-based bilstm network with lexical feature for emotion classification. In *2018 international joint conference on neural networks (ijcnn)* (pp. 1–2).

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies, 10*, 1–309.

Habimana, O., Li, Y., Li, R., gu, X., & Yu, G. (2020). *Sentiment analysis using deep learning approaches: an overview* (p. 63). Science China Information Sciences.

Hameed, Z., & Garcia-Zapirain, B. (2020). Sentiment classification using a single-layered bilstm model. *IEEE Access, 8*, 73992–74001.

Hipson, W. (2019). Using sentiment analysis to detect affect in children's and adolescents' poetry. *International Journal of Behavioral Development, 43*, 016502541983024.

Huang, X., Li, X., Zhang, L., Liu, T., Chiu, D.& Zhu, T. (2015). Topic model for identifying suicidal ideation in chinese microblog.

Ireland, D., Hassanzadeh, H.& Tran, S. (2018). Sentimental analysis for aiml-based e-health conversational agents: 25th international conference, iconip 2018, siem reap, cambodia, december 13–16, 2018, proceedings, part ii. In (pp. 41–51).

Ishaq, A., Asghar, S., & Gillani, S. A. (2020). Aspect-based sentiment analysis using a hybridized approach based on cnn and ga. *IEEE Access, 8*, 135499–135512.

Kingma, P. D., & Ba, L. J. (2015). Adam: A method for stochastic optimization. In *International conference on learning representations*.
Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence, 22*, 100–109.
Kraiss, J., ten Klooster, P., Moskowitz, J., & Bohlmeijer, E. (2020). The relationship between emotion regulation and well-being in patients with mental disorders: A meta-analysis. *Comprehensive Psychiatry, 102*, Article 152189.
Kratzwald, B., Ilic, S., Kraus, M., Feuerriegel, S., & Prendinger, H. (2018). Decision support with text-based emotion recognition: Deep learning for affective computing. *Decision Support Systems, 115*.
Lee, G., Kim, C. O., & Song, M. (2020). Semisupervised sentiment analysis method for online text reviews. *Journal of Information Science, 016555152091003*.
Li, P., Li, J., Sun, F.& Wang, P. (2017). Short text emotion analysis based on recurrent neural network. In (pp. 1-5).
Li, X., Feng, S., Wang, D., & Zhang, Y. (2019). Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems, 174*.
Lin, S., Su, W., Chien, P., Tsai, M., & Wang, C. (2020). Self-attentive sentimental sentence embedding for sentiment analysis. In *Icassp 2020–2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 1678–1682).
Lin, Y., Li, J., Yang, L., Xu, K., & Lin, H. (2020). Sentiment analysis with comparison enhanced deep neural network. *IEEE Access, 8*, 78378–78384.
Luong, M.-T., Pham, H.& Manning, C. (2015). Effective approaches to attention-based neural machine translation.
Makki, I., Alhalabi, W., & Adham, R. (2019). Using emotion analysis to define human factors of virtual reality wearables. *Procedia Computer Science, 163*, 154–164.
Mekruksavanich, S., Jitpattanakul, A.& Hnoohom, N. (2020). Negative emotion recognition using deep learning for thai language. In 2020 joint international conference on digital arts, media and technology with ecti northern section conference on electrical, electronics, computer and telecommunications engineering (ecti damt ncon) (pp. 71–74).
Oh, K., Lee, D., Ko, B., & Choi, H. (2017). A chatbot for psychiatric counseling in mental healthcare service based on emotional dialogue analysis and sentence generation. In *2017 18th ieee international conference on mobile data management (mdm)* (pp. 371–375).
Partaourides, H., Papadamou, K., Kourtellis, N., Leontiades, I., & Chatzis, S. (2020). A self-attentive emotion recognition network. In *Icassp 2020–2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7199–7203).
Pennington, J., Socher, R.& Manning, C. (2014). Glove: Global vectors for word representation. In (Vol 14, pp. 1532-1543).
Pânisoara, G., Panisoara, I., & Sandu, C. (2015). Comparative study on emotions analysis in students of psychology by gender. *Procedia - Social and Behavioral Sciences, 180*, 1638–1642.
Poornima, A., & Priya, K. S. (2020). A comparative sentiment analysis of sentence embedding using machine learning techniques. In *2020 6th international conference on advanced computing and communication systems (icaccs)* (pp. 493–496).
Rajabi, Z., Shehu, A., & Uzuner, O. (2020). A multi-channel bilstm-cnn model for multilabel emotion classification of informal text. In *2020 ieee 14th international conference on semantic computing (icsc)* (pp. 303–306).
Ren, L., Xu, B., Lin, H., Liu, X., & Yang, L. (2020). Sarcasm detection with sentiment semantics enhanced multi-level memory network. *Neurocomputing, 401*.
Rhanoui, M., Mikram, M., Yousfi, S., & Barzali, S. (2019). A cnn-bilstm model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction, 1*, 832–847.
Sailunaz, K., & Alhajj, R. (2019). Emotion and sentiment analysis from twitter text. *Journal of Computational Science, 36*.
Sathis Kumar, T., Mohamed Nabeem, P., Manoj, C. K., & Jeyachandran, K. (2020). Sentimental analysis (opinion mining) in social network by using svm algorithm. In *2020 fourth international conference on computing methodologies and communication (iccmc)* (pp. 859–865).
Senpy: A framework for semantic sentiment and emotion analysis services. (2020). Knowledge-Based Systems, 190, 105193.
Seyeditabari, A., Tabari, N.& Zadrozny, W. (2018). Emotion detection in text: a review. ArXiv, abs/1806.00674.
Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*, 427–437.
Song, M., Zhao, X., Liu, Y.& Zhao, Z. (2018). Text sentiment analysis based on convolutional neural network and bidirectional lstm model. In Q. Zhou, Q. Miao, H. Wang, W. Xie, Y. Wang& Z. Lu (Eds.), Data science (pp. 55–68). Singapore: Springer Singapore.
Song, P., Geng, C., & Li, Z. (2019). Research on text classification based on convolutional neural network. In *2019 international conference on computer network, electronic and automation (iccnea)* (pp. 229–232).
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*(56), 1929–1958.
Sun, X., & He, J. (2020). A novel approach to generate a large scale of supervised data for short text sentiment analysis. *Multimedia Tools and Applications, 79*.
Vanlalawmpuia, R., & Lalhmingliana, M. (2020). Prediction of depression in social network sites using data mining. In *2020 4th international conference on intelligent computing and control systems (iciccs)* (pp. 489–495).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.& Polosukhin, I. (2017). Attention is all you need.
Wang, P., Yan, Y., Si, Y., Zhu, G., Zhan, X., Wang, J., & Pan, R. (2020). Classification of proactive personality: Text mining based on weibo text and short-answer questions text. *IEEE Access, 8*, 97370–97382.
Wongkoblap, A., Vadillo, M., & Curcin, V. (2017). Researching mental health disorders in the era of social media: Systematic review. *Journal of Medical Internet Research, 19*, Article e224.
Wu, J., He, Y., Yu, L., & Lai, K. R. (2020). Identifying emotion labels from psychiatric social texts using a bi-directional lstm-cnn model. *IEEE Access, 8*, 66638–66646.
Xia, C., Zhao, D., Wang, J., Liu, J., & Ma, J. (2018). Icsh 2018: Lstm based sentiment analysis for patient experience narratives in e-survey tools. In H. Chen, Q. Fang, D. Zeng, & J. Wu (Eds.), *Smart health* (pp. 231–239). Cham: Springer International Publishing.
Xiao, X., Wei, P., Mao, W., & Wang, L. (2019). Context-aware multi-view attention networks for emotion cause extraction. In *2019 ieee international conference on intelligence and security informatics (isi)* (pp. 128–133).
Xie, J., Chen, B., Gu, X., Liang, F., & Xu, X. (2019). Self-attention-based bilstm model for short text fine-grained sentiment classification. *IEEE Access, 7*, 180558–180570.
XinSheng, Z., & Teng, G. (2019). Multi-head attention model for aspect level sentiment analysis. *Journal of Intelligent & Fuzzy Systems, 38*, 1–8.
Xu, D., Tian, Z., Lai, R., Kong, X., Tan, Z., & Shi, W. (2020). Deep learning based emotion analysis of microblog texts. *Information Fusion, 64*.
Yousaf, A., Umer, M., Sadiq, S., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Emotion recognition by textual tweets classification using voting classifier (lr-sgd). *IEEE Access, 9*, 6286–6295.
Zgierska, A., Rabago, D., & Miller, M. M. (2014). Impact of patient satisfaction ratings on physicians and clinical care. patient prefer adherence. Patient Prefer and Adherence, 8.
Zhang, L., Wang, S., & Liu, B. (2018). *Deep learning for sentiment analysis: A survey* (p. 8). Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
Zhang, Q., & Lu, R. (2019). A multi-attention network for aspect-level sentiment analysis. *Future Internet, 11*, 157.
Zhang, Y., Xu, B., & Zhao, T. (2020). Convolutional multi-head self-attention on memory for aspect sentiment classification. *IEEE/CAA Journal of Automatica Sinica, 7*(4), 1038–1044.
Zhang, Y., Zhao, Z., Wang, P., Li, X., Rong, L., & Song, D. (2020). Scenariosa: A dyadic conversational database for interactive sentiment analysis. *IEEE Access, 8*, 90652–90664.