



# Prediction and detection of emotional tone in online social media mental disorder groups using regression and recurrent neural networks

Sai Dheeraj Kanaparthy<sup>1</sup> · Anjali Patle<sup>1</sup> · K. Jairam Naik<sup>1</sup> 

Received: 10 May 2022 / Revised: 17 January 2023 / Accepted: 6 April 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Online social media networks have become a significant platform for persons with mental illnesses to discuss their struggles and obtain emotional and informational assistance in recent years. One such platform is Reddit, where sub-groups called ‘subreddits’ exist, based on a variety of topics including mental illnesses such as anxiety or depression. We analyse the user’s interactions to calculate the mental health status by formulating and using a parameter called ‘emotional tone’ representing the user’s emotional state. VADER sentiment analysis and TextBlob are used to categorise emotional tone and find distribution of emotional polarity and subjectivity of comments. For final tone prediction, RNN and State-Of-The-Art word embedding techniques are used to develop a predictive model. The resultant model provides end-to-end categorization and prediction of emotional tone. We obtain results with respect to Weighted L1 Loss that deals with extreme responses. The MODEL transcends all the base-lines by at least 12.1% and the final emotional status of the authors is positive.

**Keywords** Gated recurrent units · Recurrent neural networks · Sentiment analysis · Prediction · DistilBERT · VADER · TextBlob

## 1 Introduction

As per the World Health Organization’s latest update, 322 million people have been affected by depression between 2005 and 2015 and the amount of persons hurting from the mental illness increased by 18% [23]. Anxiety affects 264 million individuals, while approximately 60 million people suffer from bipolar disorder. These diseases can be a cause for suicide in

---

✉ K. Jairam Naik  
Jnaik.cse@nitrr.ac.in

Sai Dheeraj Kanaparthy  
kanaparthy.saidheeraj@gmail.com

Anjali Patle  
anjaliptle1417@gmail.com

<sup>1</sup> Department of Computer Science & Engineering, National Institute of Technology Raipur, Raipur Chhattisgarh, India

their most severe form; with roughly 800,000 suicide cases each year [22]. Most individuals who are affected are unable to receive much-needed support due to an amalgamation of limited resources, social stigma linked with mental diseases, and a fear of asking for help.

Due to the stigma around mental health and the scarcity of appropriate resources, many people suffering from diseases are unable to receive the help they need. In such a scenario, alternate media like helpline numbers, chatbots and email services have been developed to provide support to people. Social media platforms like Reddit, Facebook and Tumblr have emerged as a significant tool to help people in sharing their thoughts. One such platform Reddit is an online social networking forum which comprises 'subreddits' which are smaller subgroups and communities centered on a particular topic. A participant in the subreddit can create a post to start a thread. Other users and the post author can then reply on the post and to each other through comments. Users can also reply to the existing comments. There are several subreddits on the platform that focuses on the mental health challenges faced by users suffering from a particular disorder such as depression, anxiety and bipolar disorder [15]. Users share their problems on these subreddits and engage in interactions with other users which are generally aimed at providing encouraging and comforting words to the original poster in order to improve their state of mind positively.

In this work, we aim at analyzing the role that these user interactions can play in making a change in emotional state of a person. We look further into the distribution of emotional polarity and subjectivity of comments in subreddits to get an idea about the typical nature of interactions. We examine the user interactions to see if they can help in improving the mental health status of the author of the post by studying if the user's emotional state changes after engaging in these interactions. We also explore if this shift in emotional state can be modeled with the objective of predicting this shift in order to develop a basis for improvement in a user's mental status. This is achieved by analyzing the user interactions as a thread of comments following a post, and then modelling the emotional tone of the user as the interaction progresses.

The main objective and results of the work is as follows:

1. Collecting the data from user interactions on the mental health subreddits on Reddit platform. Pre-processing the data using TextBlob to study the emotional polarity and subjectivity of comments in the threads, in order to establish a correlation between the subjectivity of comments and the polarity of author's final comment. This analysis helps in understanding the effect of the nature of comments on the final emotional state of author. It also gives an idea about the typical nature of emotional tones used in the interactions on these sub-reddits. The outcome was that subjectivity of the comments for all the subreddits is high, which indicates that the nature of comments involved were frequently based on the user's own emotions, experience and opinions, rather than only factual information.
2. Using VADER sentiment analysis on collected raw data for categorising and assessing the emotional state of users in terms of emotional tones of each comment in the interaction.
3. Analysing the thread of comments under an author's post and establishing a correlation between the emotional tones of comments, posts and the final comment made by the author in the thread. Creating a multi-layer RNN based predictive model to forecast the last comment emotional tone of the author, after being subjected to interactions in the comments. This will help in understanding whether the user interactions affected the author positively or negatively. The final comments are mostly of positive polarity, indicating that the user interactions positively affect the author and improve their state of mind, resulting in more optimistic comments from the author.

4. Comparing various baselines and model performances and analysing them in order to produce more accuracy and efficiency. We see that the model outperforms the baselines by at least 12.1%.

We begin by studying the distribution of the emotional polarity and subjectivity of comments in the subreddit. This is done using TextBlob, a tool that is used for Sentiment Analysis, an open-source python-based library which uses Natural Language ToolKit (NLTK) to provide various NLP functionalities [20]. It produces two values: polarity and subjectivity with reference to a textual piece of data. Polarity is related to the positive or negative emotion behind a statement and it is a value between [-1, 1]. The degree of personal opinion and a fact in a text is calculated by subjectivity, which is returned as a value between [0, 1]. If a text contains more personal opinion than actual factual information, it is more subjective and less objective [5]. Here, we use these two values for the comments in the subreddit for analyzing the sentiment as well as the nature of advice and information shared in the comments.

The initial emotional tone of the user is detected using a sentiment analysis tool called VADER. Sentiment analysis uses a computational approach of subjectivity in material to quantify a speaker or writer's attitude, emotions, judgments, moods, and thoughts. VADER (Valence Aware Dictionary for Sentiment Reasoning) is a tool for analyzing sentiments, for textual material which considers both the amplitude (magnitude) and the polarity (positive/negative) of emotion [12]. The VADER sentiment analysis system employs a dictionary to link lexical features and emotional intensities, or sentiment scores. The sentiment score of a textual data can be calculated by adding the intensity of each word in the text.

The prediction model is based on Recurrent Neural Networks (RNN). It refers to a category of Artificial Neural Network (ANN) which is usually based on processing sequentially ordered data or time series. RNNs are based on 'memory' allowing them to influence current input and output with knowledge of previous inputs [13]. Here the chronologically ordered comment thread forms the series of input. It is created using Transformer architectures based on Deep Neural Networks, particularly the DistilBERT architecture. DistilBERT is a BERT-based Transformer model that is compact, fast, inexpensive, and light [17]. During the pre-training phase, knowledge distillation is used to reduce the size of a BERT model by 40%. Knowledge distillation (also known as teacher-student learning) is a compression method that involves training a smaller model to mimic the behavior of a bigger model (or an ensemble of models). This enables easier deployment of the model due to its compactness [9].

While sentiment analysis has been a popular and widely-used method in works related with emotions and opinions of people, its application in the mental health field has been somewhat limited to detection models. It has been utilized on social media platforms for recognition of a user's emotional tone; however the scope of analyzing the shift in this tone has not been explored. This study focuses on this emotional shift which provides valuable insights into the effect of the interactions on these platforms. This work develops a structure which extends beyond just categorising and identifying the mental status of an individual based on data extracted from social media, by integrating the mechanism to predict the patterns in the emotional shift.

Accurate prediction of the shift in emotional tone can be a very crucial step in formulating methodologies and tools to positively alter the mental state of a user suffering from mental health disorders. This further has scope for the development of tools that provide custom

content according to the desired shift in emotions. This forms a great scope for developing tools that can contribute to the desired shift in emotional tone such as recommendation systems, counseling bots, etc. and therefore help in early intervention in mental health issues.

The use of knowledge distillation is another promising area which will make such analysis and prediction models compact. This will enable the application of these models on time-sensitive and resource sensitive systems such as at the edge of the networks. Therefore, more compact models can be developed which can be efficiently applied on big data systems and in use cases with instantaneous inputs.

The study consists of seven sections that describe the background, importance and implementation of proposed methodology. First the introduction starts with a discussion of the emergence of social media platforms and how they cater to people with mental health issue. It also consists of a rundown of the sentiment analysis tools in use. Then, the objective, importance, scope and motivation behind the study is also discussed. Following this, Sect. 2 shares the existing literature work associated with sentiment analysis and various prediction models based on Neural Networks. The third section provides a detailed description of the modelling of the problem and also discusses the proposed architecture. Then the pre-processing block of the architecture is discussed along with a detailed description of the methodology workflow. The fourth section describes the experimental setup and technology used and the fifth section discusses the performance metrics that are used for analysis. Following this, the sixth section consists of the performance evaluation and finally the seventh section provides a conclusion of the study.

## 2 Related work

With the emergence of social media platforms and their growing popularity among the masses, their social and mental implication has been an area of great interest for researchers. Several mental health studies have been published focusing on the Reddit data, that aim at predicting the state of mental health of users through techniques of machine learning such as Support Vector Machine (SVM) and Logistic Regression and [6, 8, 18].

The authors of work [21] proposed a L2- penalized logistic regression-based model which used clustering analysis to predict the occurrence of mental issues through social media. In addition to two studies based on logical regression models, a third study using split-half analysis was integrated which focused on more recent past posts for better predictions. The authors of work [18] delved into the subject of suicide risk assessment by using CNN to introduce and exhibit the value of a novel, comprehensive rubric to assess suicide risk. They compared crowdsourced data with mastered performance, and presented baseline predictive modelling experiments using the new dataset created.

Following the research surrounding social media and mental health, The authors of work [14] used statistical learning techniques such as ERT and L2-penalized (ridge) logistic regression to build 3 models to evaluate forecasts of medical conditions from posts on social media. The first two models compared the ability to predict of Facebook language (model 1) to that of demographics (model 2). The hybrid model 3 compared incremental contribution of Facebook language to demographic.

The authors work in [16] used transfer learning classifiers that recognize the social media language indicative of mental health outcomes for studying the changes in psycho-social expressions over language and time during Covid-19. The proposed model can study not only the linguistic changes but also the temporal changes in combination with support expressions.

The authors [2] used techniques like conversation models that are sequence-based, language model comparisons, message clustering, to present a set of computational discourse analysis algorithms for calculating the relationship between various linguistic features of conversations and conversation outcomes. The novel methods developed are suited for datasets that are large-scale and can be used to discover actionable conversation strategies associated with better conversational outcomes. However, with only a small set of linguistic features, the model performs well [1].

The work proposed in [19] used Relationship Modelling Network to replicate relationship between the user and the community they interact in. Then, discussion trees were analysed, extracting descriptors that unwrap the discussed main topic. The model detects what the main messages of the set of comments are, and also gives a description of what is being discussed in different communities. This can let counsellors skip irrelevant parts and help the person in need.

The work in [11] proposed a technique which used regression machine learning models on multiple tweets and compared with outcomes of Linguistic Inquiry and Word Count Software to score tweets on a perceived anxiety scale. The model can help in determining the anxiety levels of a person based on their multiple tweets with very small error. The work authors in [3] used Machine learning models to classify short comments into positive and negative data using character level analysis to find precision, and accuracy of the model. It is predicted that classifying with a short statement uttered daily will lead to early identification of mental problems. However, the accuracy of classification can be bettered by using words' vector representation that includes their meaning.

The work proposed in [7] employed logistic regression framework to distinguish between those who have suicide intent vs. those who have shifted from a mental illness state to suicide state. This study established a method that tells if a person engaged in mental health discussions is prone to shift to suicidal ideation.

Previous works (Table 1) include classification of mental health status of users using Support Vector Machine and Recurrent Neural Network, where the column for performance attainment corresponds to the accuracy to the models proposed. Many models and architectures have been developed in order to create more accurate systems. Architectures like DistilBERT aim at creating compact version of these technologies in order to improve their ability to be deployed at network edges. This forms a solid groundwork for applying the sentiment analysis methods in prediction models to create a more robust system where classification of emotions can be extended to analysis the influence of various factors and model predictions based on it.

The proposed architecture aims at solving the following gaps:

1. The existing approaches refined and classified the data based on a score value to predict the mental state, using methods such as Support Vector Machines, Logistic Regression, and Deep Learning [6, 8, 18]. This is in contrast to the approach followed in this paper, which is forecasting social interactions effect as a regression-based task.
2. There exists a scope of studying the continuously changing emotional tone in online mental illness group interactions in order to predict their effects. This work aims at creating such a prediction model by modelling the user's mental state during his/her interaction in a sequence of comments to forecast shifts in their emotional tone, to get a more accurate result. Analysis of the shift in the mental state of author after each comment is explored, and the determination of final mental state of author in relation to final comment is predicted.

**Table 1** Table of related work

Sno.	Author and Year of publication	Work description	Performance Attainment (%)	Model expertise & rip-offs
1	[21], (2019)	Using L2- penalized logistic regression models along with clustering analysis to predict future mental illness from social media.	81	In addition to two studies based on logical regression models, a third study using split-half analysis is integrated which focuses on more recent past posts for better predictions. The predictive ability of the model can be undermined as few of the top predictors learned by the future prediction model were explicit references to medications and therapy, as supposed to linguistic suggestions.
2	[18], (2018)	Using CNN to introduce and demonstrate the value of a new, detailed rubric for assessing suicide risk, compare crowd-sourced with expert performance, and present baseline predictive modeling experiments using the a new dataset created.	67	The new dataset created includes expert ratings for 245 users and crowdsourced ratings for a superset of 865 users which can be used for risk assessment for suicidality based on social media. The four-way classification used is just an initial exploration of the problem and hierarchical classification should be explored to cut through noise in the dataset.
3	[14], (2019)	Used statistical learning techniques such as ERT and L2-penalized (ridge) logistic regression to build 3 models to evaluate the predictability of medical conditions from social media posts	79	The first two models compares the predictive ability of Facebook language (model 1) to that of demographics (model 2). The hybrid model 3 compares incremental contribution of Facebook language to demographic. Logistic regression and extremely randomized trees as modeling algorithms are used but prior work has demonstrated, transfer learning from pre-trained text-based multilayer neural network architectures could lead to higher predictive performance.
4	[16], (2020)	Using transfer learning classifiers that identify the social media language indicative of mental health outcomes for examining the changes in psychosocial expressions over time and language during Covid-19.	97	The proposed model can study not only the linguistic changes but also the temporal changes in combination with support expressions. Twitter API used does not provide enough sample data.
5	[1], (2016)	Used techniques like sequence-based conversation models, language model comparisons, message clustering, to propose a set of computational discourse analysis methods to measure how various linguistic aspects of conversations are correlated with conversation outcomes.	62.5	The novel computational discourse analysis methods developed are suited for large-scale datasets and can be used to discover actionable conversation strategies that are associated with better conversational outcomes. The model performs well with only a small set of linguistic features.

Table 1 (continued)

Sno.	Author and Year of publication	Work description	Performance Attainment (%)	Model expertise & rip-offs
6	[19], (2018)	Used Relationship Modeling Network to model the relationship between the user and the community in which they interact. Then, discussion trees are analyzed, extracting descriptors that uncover the main topic discussed.	73	The model detects what the main messages of the set of comments are, and also gives a description of what is being discussed in different communities. This can let counselors skip irrelevant parts and help the person in need. The study does not include the type of people that tend to not take the community seriously and the irrelevant comments that come with it. The model analyses large quantities of counselor conversations and lays out strategies that can lead to better outcomes. The model does not work for real-time systems and works as a training system instead.
7	[2], (2016)	Used an unsupervised conversation model - a Hidden Markov Model to find the strategies used to lead to positive conversational outcomes.	80	Helps to determine the anxiety levels of a person based on their multiple tweets with very small error. The ratings provided by our zero acquaintances, do not actually measure felt anxiety by the tweeter, but rather perceived anxiety.
8	[11], (2019)	Used regression machine learning models on multiple tweets and compared with outcomes of Linguistic Inquiry and Word Count software's to score tweets on a perceived anxiety scale.	67	By being able to classify with a short comment that is said on a daily basis, it can be expected to lead to early detection of mental disorders. The accuracy of classification can be improved by using vector representation of words that includes their meaning.
9	[3], (2020)	Used Machine learning models to classify short comments into positive and negative data using character level analysis to find precision, and accuracy of the model.	81	A method is established that tells if a person engaged in mental health discussions is prone to shift to suicidal ideation. Throwaway/multiple accounts can be created that can shift from Mental Health sub-reddit to Suicide Watch sub-reddit, which will not be captured.
10	[7], (2016)	Used logistic regression framework to distinguish between those who have suicide intent vs. those who have shifted from a mental illness state to suicide state.	71	

### 3 Shifts modelling in emotional tone

#### 3.1 Polarity and subjectivity of comments

The TextBlob library is an open-source python Natural Language Processing (NLP) tool which can perform sentiment analysis on text using lexicon-based approach. It takes the input text and generate two values one for polarity and one for subjectivity of the text. This gives us the sentiment type i.e., positive or negative sentiment behind a comment. It also gives us the level of objectivity in a sentence and hence provides the degree of personal opinions and emotions involved in a particular comment. Thus, nature of the comments in a subreddit can be analyzed in terms of these two factors. Exploring these factors gives us an idea about the prevalent emotional state and temperament of comments in these subreddits, which can be correlated to the effect of interactions on the post author. We also specifically focus on the polarity of the last comments in all threads to see the final emotional status of the author.

#### 3.2 Emotional tone

In a particular publication, that can be either a post or a document, the emotional intensity or tone is measured by a sentiment analysis tool called VADER. This is a tool built from Open Data, and calculates four values from a piece of text provided: positive, negative, neutral and compound. We focus here on Compound, as it represents the normalization of three variables between  $-1$  (extremely negative) to  $+1$  (extremely positive). This is the value that depicts the emotional tone.

#### 3.3 Prediction task

We take the post  $q$  created by the author as  $b$  that starts a thread in the community. This thread consists of a series of comments  $d_1, d_2, d_3, \dots$  which are sorted chronologically. The last comment will be  $d_n$  made by author  $b$ . This forms a sequence  $T = \{q, d_1, d_2, \dots, d_{n-1}\}$ . Here, the goal is to determine the emotional tone,  $ET(d_n)$  of the author's last comment and of the post  $ET(.)$ , in order to understand how the author's emotions are affected by the comments in thread  $T$ . This way, the emotional impact and the help given to the members of the community can be measured, to judge how well they're doing. Posts and comments are used as input for the models, to predict the emotional tones and their shifts.

#### 3.4 Model overview

The model takes the ordered input  $T = \{q, d_1, d_2, \dots, d_{n-1}\}$  where  $q$  refers to the original post and  $d_1, \dots, d_{n-1}$  are the comments present in the thread. The last comment  $d_n$  made by the publisher is not included in this thread. The pre-processing block takes each comment  $d_i$  and extracts the following:

- i) TextBlob polarity and subjectivity score.
- ii) DistilBERT embedding.
- iii) VADER emotional tones.
- iv) Variable that shows if the comment  $d_n$  was by the post-author.



The embeddings obtained are next passed through a Gated Recurrent Unit (GRU). In this model we use a 2-layer bidirectional RNN as shown in Fig. 1. Hidden states are joined and after passing through the 2nd Fully Connected (FC) layer, the output is produced. The output is the prediction  $y$  which refers to the last comment's emotional tone  $d_n$ . Figure 1 shows the proposed architecture.

### 3.5 Limitations

One of the limitations of using Reddit posts as dataset is that it is not possible to examine if there was an actual positive or negative change in the mental health status of the user, as all the users on the platform are anonymous. It is also not possible to monitor the effect of other user interactions in the community, such as viewing and upvoting/downvoting other posts.

Another limitation is the miscalculation of emotional tones in the sentiment analysis process. This can happen for texts containing ambivalent tones, which makes it difficult to accurately classify the emotional tone of the message. Finally, it is not possible to take into account any additional help, interactions or factors in the user's life that may positively or negatively his/her mental state.

### 3.6 Pre-processing

The model is trained using word-embedding methods, as the prediction task is to happen as a part of real-time applications where embeddings serve the purpose. These embeddings are obtained using the DistilBERT architecture. Every comment  $n \in T$ , goes through DistilBERT to give an embedding " $e_n$ ". With this, we combine emotional tone as determined earlier  $ET(n)$  and a variable that displays if the author of  $n$  is the post author. Together, this is put through a fully connected (FC) layer to flatten the dimensions of embeddings that finally forms the feature vector  $Y_n$ , as described in Eq. 1.

$$Y_n = [FC(e_n), E_m T(n), \text{if } author\_post(n)] \quad (1)$$

Following pre-processing each sequence  $T$ , a feature matrix is formed, given by Eq. 2.

$$Y = [Y_q; Y_{d1}; Y_{d2}; \dots Y_{d_{n-1}}] \text{ and scalar } y = E_m T(d_n) \quad (2)$$

Each element of the feature matrix  $Y$  is put through Gated Recurrent Units that can be either forward propagated or backward only or a combination of both, making it bi-directional. The output of these units are then put through a Fully connected layer to produce the prediction vector  $y = E_m T(d_n)$ . Figure 2 describes the methodology used.

### 3.7 Datasets

A python scrapper is used which takes the data returned by Reddit API from the four sub-reddits namely *Anxiety*, *Depression*, *Bipolar* and *SuicideWatch*. Various posts are extracted from these communities and converted into a thread of post and comments (Fig. 2).

### 3.8 Pseudo code

#### 1) Scrapper

- 1.1) Collect Posts from 4 subreddits [Depression, Bipolar, Anxiety, SuicideWatch] using *fetch(author, id, num of comments, asc sort)* and store it in dictionary format, in "*post2data.pkl*".
- 1.2) Collect the comment Ids for the particular posts using post ids and *fetch()* function. Store the obtained comment Ids in "*post2comments.pkl*".
- 1.3) Iterate through "*post2comments.pkl*" to get all the comment ids. Using these comment ids, collect comments using *fetch()* and store it in "*comment2data.pkl*".

This creates the collection of threads, given by:

$$T = \{q, d_1, d_2, \dots, d_{n-1}\}$$

#### 2) Pre- Processing Block

- 2.1) Create folders for results, model, data
- 2.2) Import libraries - pandas, numpy, pickle, time, torch etc.
- 2.3) Load *DistilBERT* to convert comments into embedding ( $e_m$ ).
- 2.4) Load *TextBlob* and produce sentiment polarity and subjectivity score.
- 2.5) Load *VADER* for sentiment analysis.
- 2.6) Pass embeddings through *VADER* to give a value between  $-1 < 0 < 1$   $E_m T(m)$ .

Following pre-processing each sequence  $T$ , a feature matrix is formed, given by:

$$Y = [Y_q; Y_{d1}; Y_{d2}; \dots Y_{dn-1}] \text{ and scalar } y = E_m T(d_n)$$

#### 3) Training the model

- 3.1) Build the source data in order to forecast the  $E_m T$  based on the comments sequence in a thread
- 3.2) Create dataset using *createTensorDataset()* function
- 3.3) Get indices of the threads to be used as a dataset, using *split\_indices()* function
- 3.4) Load Training, validation, test sets using *Dataloader()* and *Subset()* function
- 3.5) Train the model using the *stacked Bi-GRU Model*, using 20 epochs and show weighted l1 loss

#### 4) Generate results table and plot graphs

## 4 Emotional tone shifts characterization

In order to understand emotional tone shift of users within a user interaction, we select segments of threads with the author of a post present in which:

1. The author starts the post and ends with their last comment before the author moves into another thread,
2. Has minimum one comment from a different user, and
3. Has within 24 h between the next posts.

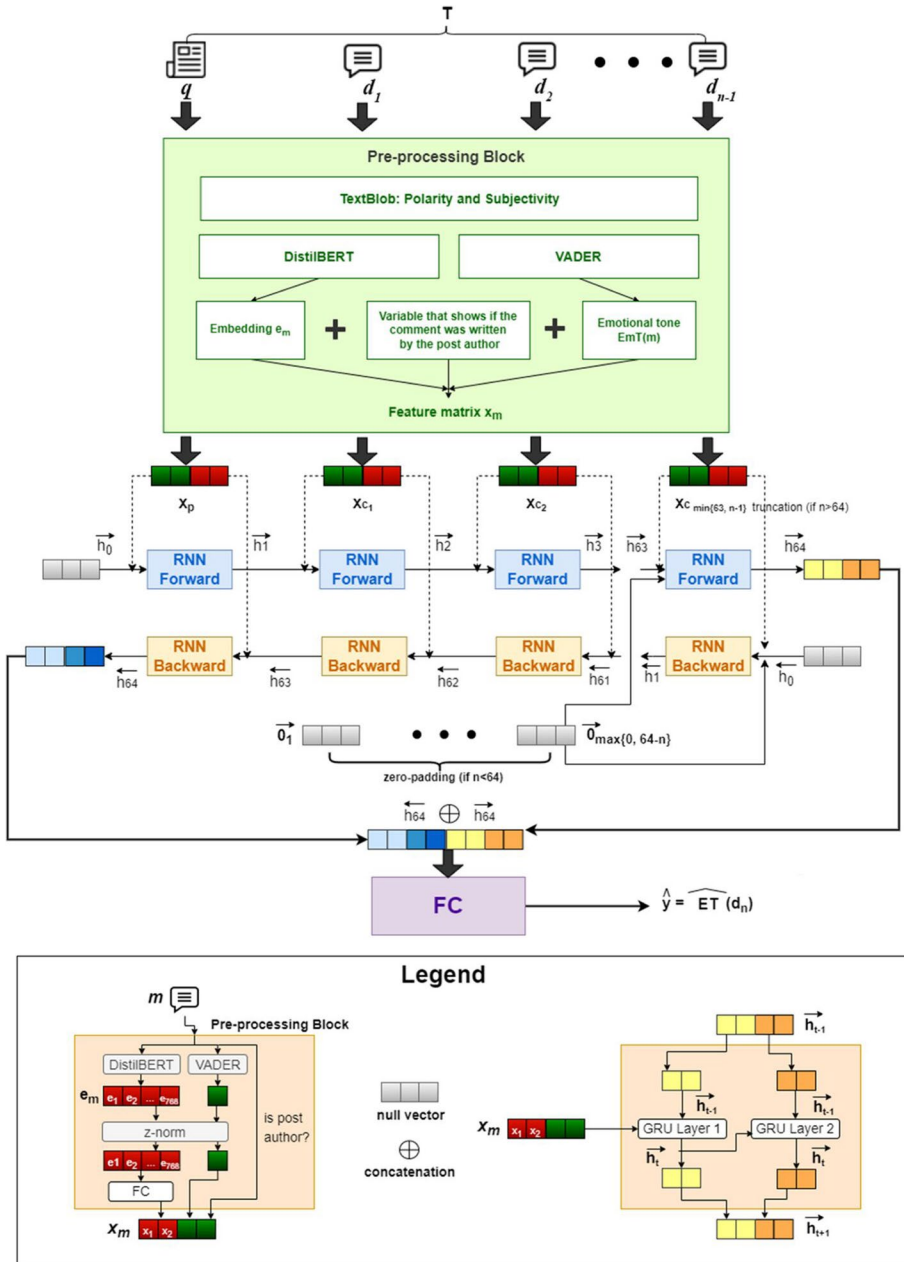


Fig. 1 Proposed architecture

Figure 3 shows these criteria followed for the thread selection process.

We take 24 h into consideration, as the longer duration might cause a different effect on the author’s emotional state. We also take threads that start with the post and ends with the last comment of the author despite having more comments within the thread, to create a

Fig. 2 Flowchart

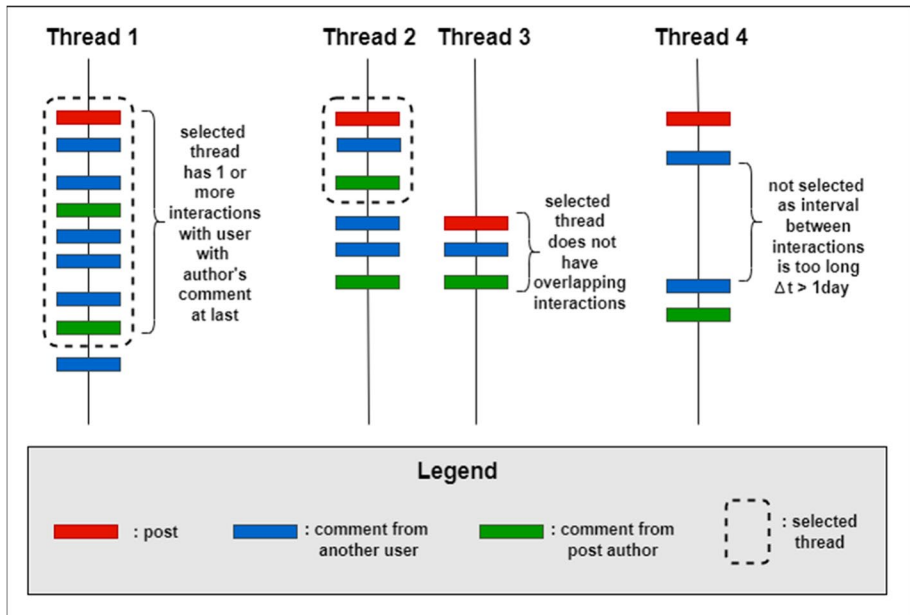
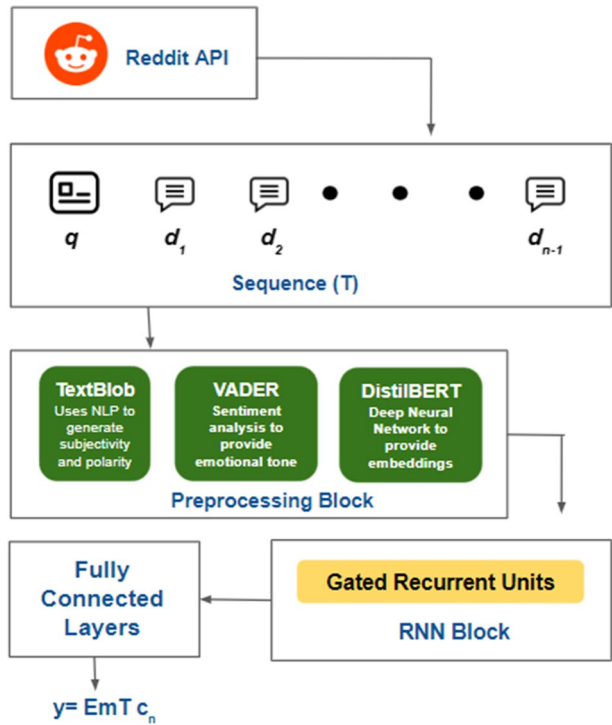


Fig. 3 An example of thread selection process

prediction. Threads that overlap i.e. threads created at the same time or a part of an existing thread being extended and at the same time a new thread is created, the part that is different is taken, and the rest is truncated. The unique part is considered.

Next, we move towards finding the emotional tone within the threads. We consider the emotional tone of:

1. The post,
2. The last comment made by the author,
3. The average emotional tone of each comment thread (based on comments of other users).

The posts put in these subreddits (Depression, Bipolar, Anxiety, and SuicideWatch) are generally depictions of their traumas, which results in producing a negative emotional tone and negative polarity.

The subjectivity of the comments in these subreddits is also high as users often share their personal emotions, experiences and advice in order to provide comfort to the post authors seeking help.

The last comment made by the author in a thread is used to determine his emotional tone. From experimentation, we observe a positive higher value obtained of the emotional tone, which suggests that the supportive comments made by other users have helped the author to feel better as a result.

The average emotional tone generally pans out to be positive, as supportive comments are offered to the author, thus resulting in a positive value for the emotional tone.

Overall, the results display the fact that being a part of respective mental illness groups, the familiar environment provides empathy and support for each other, thus making an improvement in one's emotional being.

## 5 Experimental setup

The execution and performance analysis of the proposed model was conducted using Jupyter Notebook software. The core algorithm was developed using Python and compiled using Python 3.3.2.

The experiments were conducted using a HP PC consisting of an Intel® Core™ i7-7300HQ 2.50 GHz CPU in a 64-bit Windows-10 OS.

To obtain the predictions of emotional tone, from the sequence of data extracted from different subreddits, we use RNN as it performs well with sequential data. We have used RNN using Gated Recurrent Units (GRUs), as GRUs help making a decision as to what information is to be passed on to the output as compared to a standard RNN. Since we're finding the prediction of emotional tone in a numeric form, implying a regression task, the hidden layer of RNN (half of  $Y_m$ ) is passed through a Fully Connected (FC) layer to get the scalar prediction  $y$  for  $EmT(c_n)$ . If a unidirectional RNN is used, hidden state of the last comment in sequence T is used, whereas a bi-directional RNN, the final hidden states are joined i.e. the forward and backward halves of the network.

**Loss function** Since this is a regression task, mean squared error (MSE Loss) and mean absolute error (L1 loss) are generally used as loss functions but with training using these, the model was unable to gather predictions along the average, thus preventing from accurate predictions of further extreme values of emotional tone  $EmT(c_n)$ . Hence, we use

Weighted L1 loss because the distribution of emotional tone  $y$  varies between  $-1$  and  $1$ , forming a bell curve [4]. We weight the absolute error for  $(X, y)$  proportionally to the reciprocal of the number of elements in its bin. 10 equal-length bins are used [10].

**Train/validation/test sets and training** From the data extracted from each subreddit, namely: Depression, Bipolar, SuicideWatch and Anxiety, we perform a random split of 80-10-10%, stratified by weight bins to create train, validation and test set for the model.

The model is designed to yield minimum validation losses, and is trained from the entire dataset (MODEL) using batch size 32 and optimizer Adam over 20 epochs. We find that the training generally ends within 8 epochs.

## 6 Prediction evaluation

**Weighted L1 loss** A better loss function in which, since the distribution of observations is a bell curve, observations are binned w.r.t  $y$  and the absolute error is weighted for  $(Y, y)$  proportionally to the reciprocal of the number of elements in its bin [4].

The proposed model is evaluated using the following baselines:

1. *Mean*: This baseline is a heuristic that gives the overall average emotional tone in sequence of comments  $T$ .

$$\widehat{y}_{\text{MEAN}} = 1/|T| \sum_{m \in S} \text{EmT}(m) \quad (3)$$

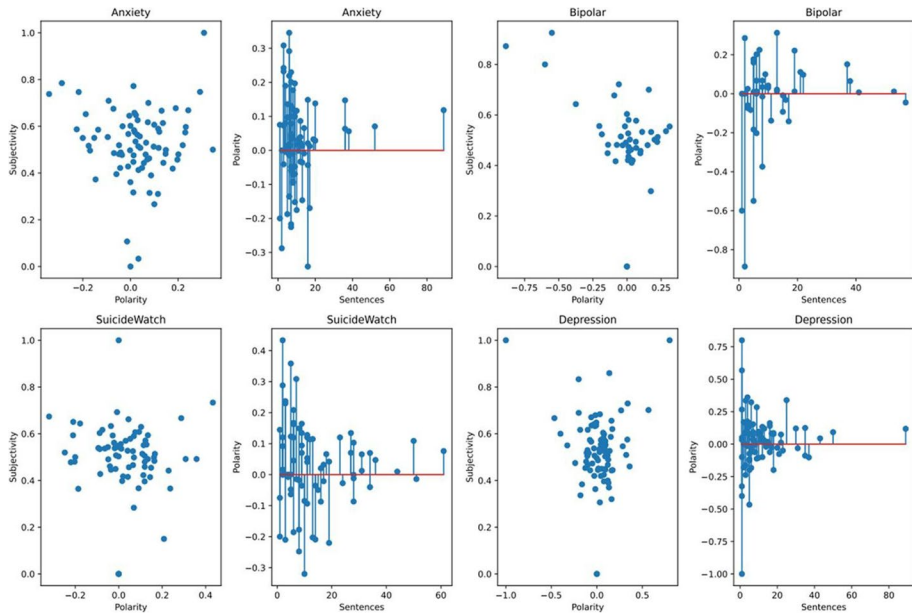
2. *Last*: a heuristic baseline that gives the last comment's emotional tone in  $S$ .

$$\widehat{y}_{\text{LAST}} = \text{EmT}(c_{n-1}) \quad (4)$$

### 6.1 Qualitative results

First we analyze the results of TextBlob sentiment analysis which gives us the information about the sentiment polarity and degree of subjectivity of the text. In Fig. 4, the figures in the first row under each subreddit shows the distribution of polarity and subjectivity of all the comments of all the threads combined in the particular subreddit. From these figures, we can see that the subjectivity of the comments for all the subreddits is high, which indicates that the nature of comments involved in user interactions is mostly subjective and frequently based on the user's own emotions, experience and opinions, rather than only factual information.

The figures in the second row under each subreddit shows the magnitude of polarity vs. the number of sentences for all the comments in a particular subreddit. From the figure it is evident that the polarity value of comments is a good mix between negative and positive quantities, with a slight inclination towards positive polarity. This is because the comments under the posts on these subreddits can be of both positive and negative emotion as the users are generally in a negative state of mind due to these mental health issues.



**Fig. 4** Distribution of polarity and subjectivity, polarity vs. the no. of sentences for each subreddit: anxiety, bipolar, suicidewatch and depression

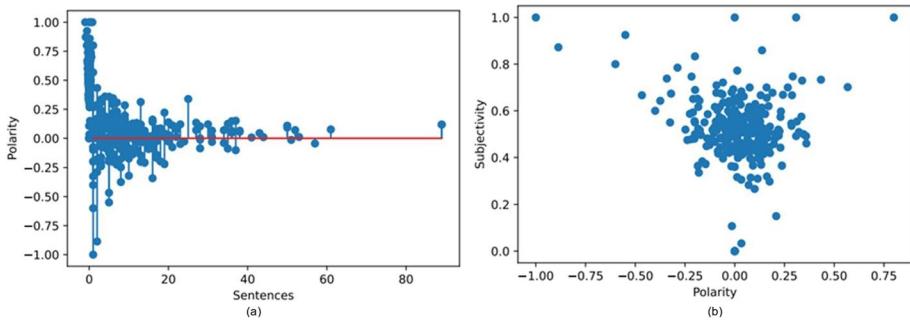
However, the slight positive inclination shows that the users share positive comments more frequently in order to help and support each other.

Figure 5a shows the combined results of polarity vs. number of sentences for all the subreddits. It can be seen that though there are significant number of sentences with negative polarity, the general inclination is towards positive interactions. Figure 5b shows the distribution of subjectivity and polarity for all the subreddits combined, and it can be inferred that for all subreddits, the user interactions are more subjective than objective in nature.

For studying the sentiment of the author's final comments, we analyze the polarity values obtained for the author's last comment for every thread in each subreddit. Figure 6 shows the results on combining these values for each subreddit. We can see that the final comments are mostly of positive polarity, indicating that the user interactions positively affect the author and improve their state of mind, resulting in more optimistic comments from the author.

Figure 7 shows the polarity vs. sentences for last comment on all the threads of all the subreddits combined. This also provides us the same inference that though there exist comments with negative polarity, maximum comments show a positive magnitude showing that the shift in emotional tone of the author is a positive one.

The results in all the graphs above give us a qualitative overview of the effect of user interactions on the author's emotional tone. The overall positive inclination of the comments indicates that user interactions on these subreddits are helpful in positively influencing the user's state of mind. The subjectivity analysis of the comments also gives us an idea about the nature of the comments on these platforms and shows that most of the comments are subjective, coming from user's own opinions and emotions, rather than factual information. Next, we analyze the predictive modeling of this emotional shift.



**Fig. 5** **a** Polarity vs. no. of sentences in all subreddits combined **b** Distribution of Polarity and Subjectivity of all comments in all subreddits combined

We plot the joint distribution for the values of true and predicted magnitudes of the MODEL against two other baselines as shown in the heat map in Fig. 8. 100 random threads from the test set as well as a Gaussian Kernel Density estimator which has been color-coded, represented by circles. The sizes of the circles are proportional to the number of comments, which lie in the range of 2 to 64. The MEAN baseline outputs the average emotional tone in sequence  $S$ , and MEANS predictions are highly concentrated around  $[-0.3, 0.8]$ .

The LAST baseline outputs the emotional tone of the last comment in  $S$  which contains a large range of prediction, but it is observed that it is concentrated along the diagonal in the Depression subreddit. The larger the vertical difference to the diagonal, more errors is produced, because of extremely negative responses within the subreddit. We see here in Fig. 8 that the extreme values are best predicted by MODEL.

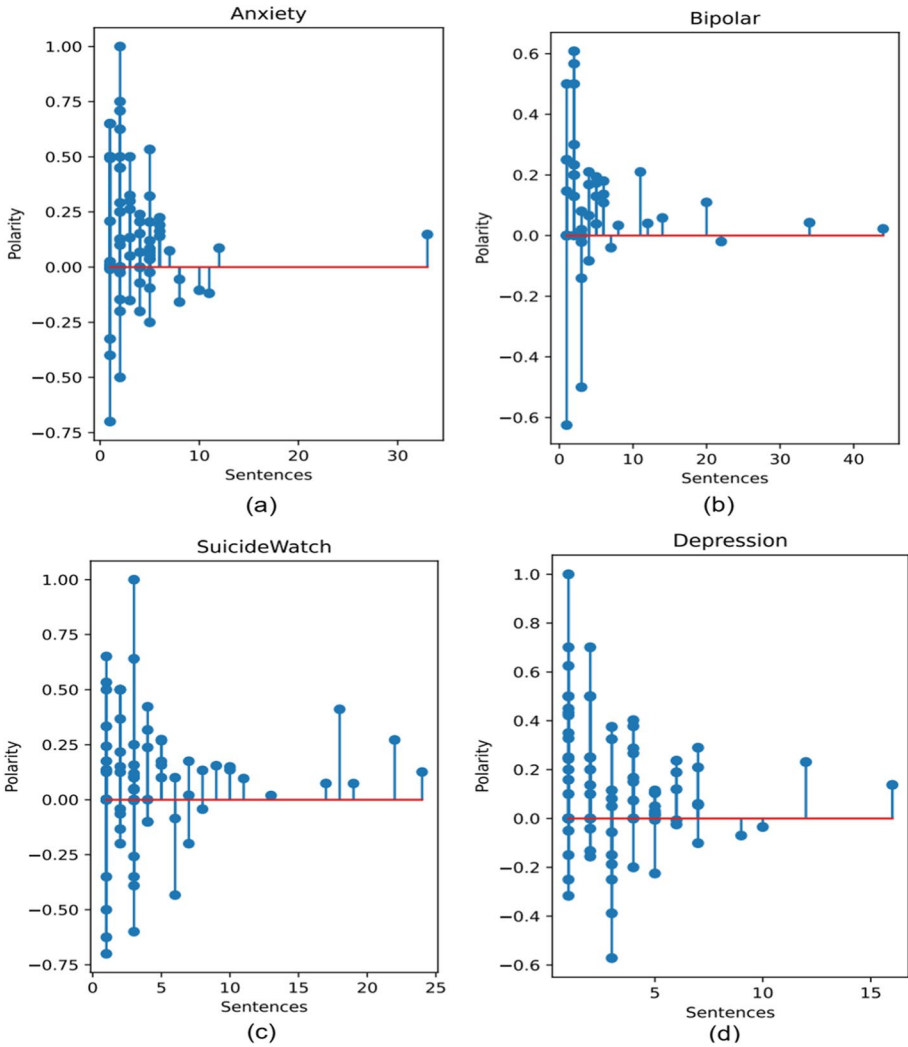
We also use XGBoost, which is a gradient boosting bundle of decision trees, which perform with high performance and deliver accurate results in prediction tasks. XGB is used on Train, Test and Validation datasets, to see if the desired results are being achieved on the Test dataset after working on the Validation dataset. We observe that the TEST dataset outperforms the Validation dataset, which ensures that the proposed model is trained well, as shown in Fig. 9.

## 6.2 Quantitative results

We abbreviate each subreddit as follows: Anxiety as ANX, Depression as DEP, Suicide-Watch as SUI and Bipolar as BIP for simplicity. Comparing various baselines in order to produce more accuracy and efficiency, we obtain results in terms of the Weighted L1 Loss that deals with extreme responses, as shown in Table 2. The MODEL performs better than all the other baselines by at least 12.1%.

**Prediction examples** We take an example thread from the subreddit *Depression*, where the model has accurately predicted the final emotional tone 0.43 for the thread author's final emotional tone 0.33. This also shows how the emotional tone has improved from  $-0.11$  to 0.42 on receiving positive feedback and support from their peers. We can clearly observe the same from the Table 3 below.





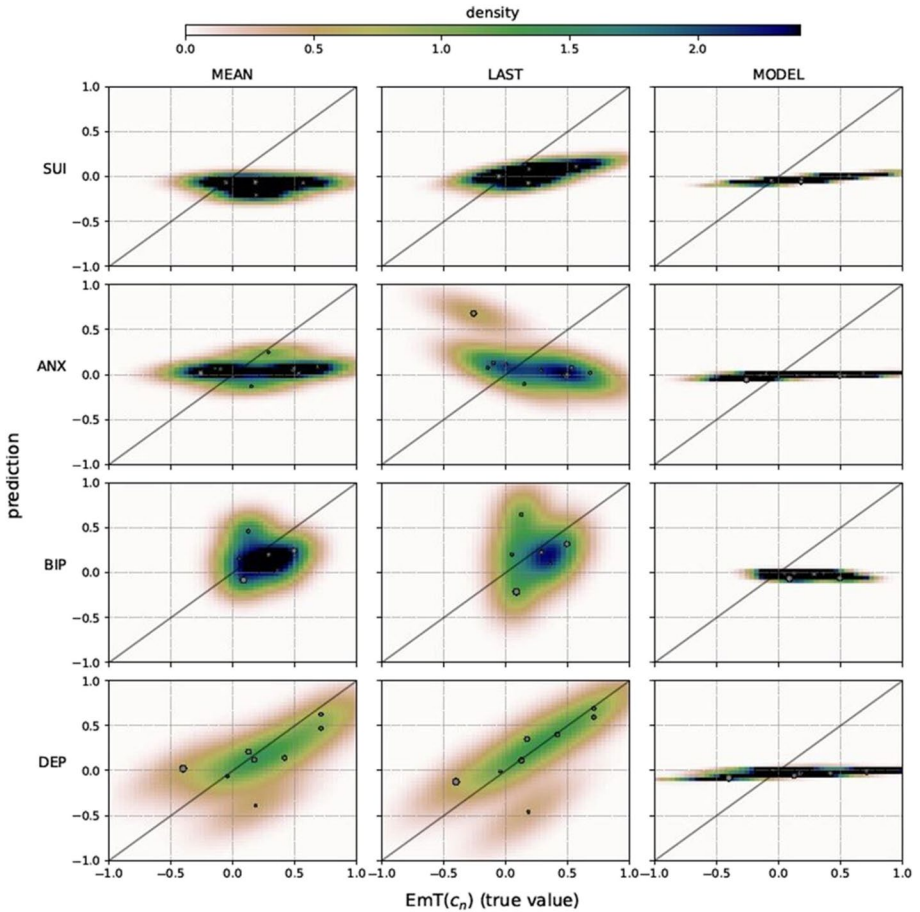
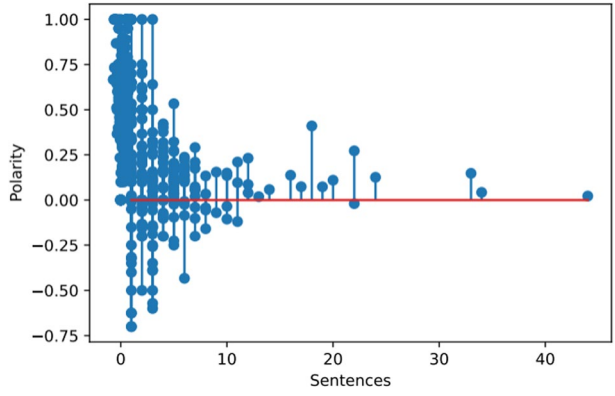
**Fig. 6** Polarity vs. no. of sentences for author’s last comment in each subreddit **a** Anxiety **b** Bipolar **c** SuicideWatch and **d** Depression

## 7 Conclusions

We address the topic of predicting progression in users’ emotional tone (a substitute for emotional state) following conversations in mental disease online forums in this work.

Sentiment analysis using TextBlob and VADER shows that the final emotional status of the authors is rendered positive after being subjected to interactions on the subreddits. The results of our prediction model showed that, despite the task’s inherent difficulty, a possibility exists, to predict how users will react to interactions on these mediums. The qualitative result analysis shows that the TEST dataset outperforms the validation dataset

**Fig. 7** Polarity vs. no. of sentences for author's last comment all subreddits combined



**Fig. 8** Heat Map of joint distribution of true and predicted values

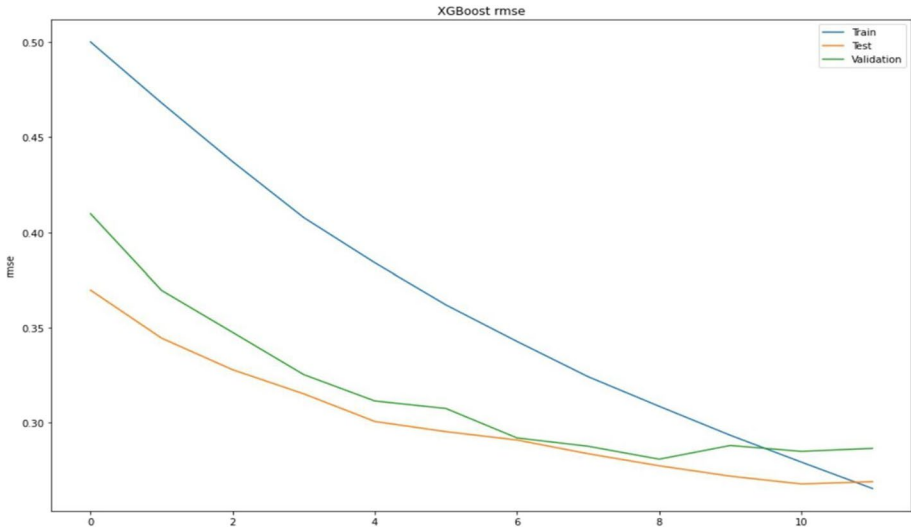


Fig. 9 Comparison of performance between train, test and validation dataset using XGBoost.

Table 2 Result with respect to weighted L1 loss

Predictor	Weighted L1 loss			
	ANX	BIP	DEP	SUI
MEAN	0.491	0.191	0.486	0.355
LAST	0.654	0.182	0.288	0.245
MODEL	0.392	0.173	0.217	0.211

Table 3 Example of user interaction and subsequent emotional tone progression

Post and comments	EmT
Author: I want to die. I hate everything. It's greedy of anyone who disagrees with suicide.	-0.11
Commenter: A suicide is usually not the way out, and that's the thought of neurotypical people, however people like us with severe anxiety or depression - do not agree with that thought.	0.69
Author: Thank you, glad to know we're in this sinking boat together.	0.14
Commenter: We are, and I may never meet you, but suicide is never the answer. I was at the verge too, but trust me, it gets way better. You just have to wait and let things run its course. You have so much potential, and it would be the entire humanity's loss if you leave.	-0.09
Author: I feel way better, thank you for your kind words. Happy to know that the world is not that bad as I perceive it to be.	0.42

and hence it indicates that the model is well-trained. The quantitative analysis shows that the MODEL transcends the Mean and Last baselines by at least 12.1%.

The models we're considering don't need any personally identifiable information specific to a user, so we can deploy these models while avoiding any privacy concerns. They are also advantageous and relevant for newcomers to a community. New systems can be employed which can make correct predictions to help users in writing comments, increasing their chances of having a good impact on their readership. Furthermore, the prediction model can be improved by integrating attention-based models, RNNs, Transformers and Graph Neural Network models.

**Data availability** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## Declarations

**Competing interests** All authors have participated in (a) conception and design, analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version. This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue. The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

**Conflict of interest** The authors declare no potential conflict of interest.

## References

1. Althoff T, Clark K, Leskovec J (2016) Natural language processing for mental health: large scale discourse analysis of Counseling Conversations. ArXivabs/1605.04462
2. Althoff T, Clark K, Leskovec J (2016) Large-scale analysis of counseling conversations: an application of natural language processing to mental health. *Trans Assoc Comput Linguist*
3. Baba T, Baba K, Ikeda D (2020) Detecting Mental Health Illness Using Short Comments. In: Takizawa M, Xhafa F, Barolli L, Enokido T, (eds), *Advanced Information Networking and Applications - Proceedings of the 33rd International Conference on Advanced Information Networking and Applications AINA-2019*. Springer Verlag.
4. Bárbara Silveira HS, Silva F, Murai, da Silva APC (2021) Predicting user emotional tone in mental disorder online communities. *Future Gener Comput Syst* 125:641–651
5. Caplan S, Alvidrez J, Paris M, Escobar JI, Dixon JK, Desai MM, Whittemore R, Scahill LD (2010) Subjective versus objective: an exploratory analysis of Latino primary care patients with self-perceived depression who do not fulfill primary care evaluation of mental disorders patient health questionnaire criteria for depression. *Prim Care Companion J Clin Psychiatry* 12(5):PCC.09m00899
6. Chancellor S, Hu A, De Choudhury M (2018) Norms matter: Contrasting social support around behavior change in online weight loss communities. In: *SIGCHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, pp 1–14
7. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M (2016) Discovering shifts to suicidal ideation from Mental Health Content in Social Media. *Proc SIGCHI Conf Hum Factor Comput Syst*
8. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M (2016) Discovering shifts to suicidal ideation from mental health content in social media. In: *SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp 2098–2110
9. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, pp 4171–4186. <https://aclanthology.org/N19-1423>. Accessed 15 Apr 2022

10. Gkotsis G, Oellrich A, Velupillai S et al (2017) Characterisation of mental health conditions in social media using informed deep learning. *Sci Rep* 7:45141
11. Gruda J, Hasan S (2019) Feeling anxious? Perceiving anxiety in tweets using machine learning
12. Hutto C, Gilbert E (2014) VADER: a parsimonious rule-based model for sentiment analysis of social media text. *Proc Int AAAI Conf Web Soc Media* 8(1):216–225. Retrieved from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>. Accessed 3 Apr 2022
13. Julia Ive G, Gkotsis R, Dutta R, Stewart, Velupillai S (2018) Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. LA. Association for Computational Linguistics, New Orleans, pp 69–77. <https://aclanthology.org/W18-0607>. Accessed 23 Mar 2022
14. Merchant RM, Asch DA, Crutchley P, Ungar LH, Guntuku SC, Eichstaedt JC, Hill S, Padrez K, Smith RJ, Schwartz HA (2019) Evaluating the predictability of medical conditions from social media posts. *PLoS One* 14(6):e0215476
15. Park A, Conway M (2018) Harnessing reddit to understand the written communication challenges experienced by individuals with mental health disorders: analysis of texts from mental health communities. *J Med Internet Res* 20(4):e8219
16. Saha K, Torous J, Caine ED, De Choudhury M (2020) Social media reveals psychosocial effects of the covid-19 pandemic, *MedRxiv*
17. Sanh V, Debut L, Chaumond J, Wolf T (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. <https://doi.org/10.48550/arXiv.1910.01108>
18. Shing H-C, Nair S, Zirikly A, Friedenber M, Daumé H III, Resnik P (2018) Expert, crowdsourced, and machine assessment of suicide risk via online postings. In: *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. Association for Computational Linguistics, New Orleans, pp 25–36. <https://aclanthology.org/W18-0603>. Accessed 16 Mar 2022
19. Silveira Fraga B, Couto da Silva AP, Murai F (2018) Online social networks in health care: a study of mental disorders on reddit. *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*
20. TextBlob Simplified Text Processing-. <https://textblob.readthedocs.io/en/dev/#textblob-simplified-text-processing>. Accessed 26 Apr 2022
21. Thorstad R, Wolff P (2019) Predicting future mental illness from social media: a big-data approach. *Behav Res Methods* 51(4):1586–1600
22. WHO (2017) Depression and other common mental disorders: Global health estimates, Tech. Rep., World Health Organization. <https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf?sequence=1&isAllowed=y> Accessed 5 Feb 2022
23. WHO, Mental health action plan 2013–2020 (2020) <https://www.who.int/publications/i/item/9789241506021>. Accessed 8 Mar 2022

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.