Social Network Analysis and Mining (2023) 13:141
https://doi.org/10.1007/s13278-023-01145-1

ORIGINAL ARTICLE

Identifying discernible indications of psychological well-being using ML: explainable AI in reddit social media interactions

Pahalage Dona Thushari[1] · Nitisha Aggarwal[2] · Vajratiya Vajrobol[2] · Geetika Jain Saxena[3] · Sanjeev Singh[2] · Amit Pundir[3]

Received: 9 May 2023 / Revised: 25 September 2023 / Accepted: 26 September 2023 / Published online: 25 October 2023

Abstract

Psychological well-being is a multidimensional construct and identifying it using a systematic, comprehensive approach offers insights fundamental to critical outcomes. Social networks are valuable resources for research, providing a pragmatic way of generating empirical evidence on psychological well-being based on the textual indicators across different populations. This study analyzed the information on various Reddit social media groups dedicated to mental health. The classes, namely depression, anxiety, bipolar, and SuicideWatch, using the SWMH dataset have been analyzed. The text-based interactions of persons with mental illness have common motifs like negative language and expressions like 'hopelessness,' 'emptiness,' or 'helplessness.' Topic modeling identified recurring themes and subjects that helped classify discernible factors influencing mental health. Classifiers for multiclass classification to classify targeted mental health issues based on users' network behavior and posts were trained and tested to get predictions on context (e.g., MentalBERT) and non-context-based (e.g., LR and NB) models. The MentalBERT model outperformed the other eight baseline models with an average accuracy of 76.70%, which is 4% more than reported in previous studies. Explainable AI was used to examine the trustworthiness of each model, and the explanations were evaluated using the LIME model. Explainability is crucial as mental health data characterizes syndromes, outcomes, disorders, and signs/symptoms exhibiting probabilistic interrelationships with each other. Explanations of these intricate interconnections can assist the extensive research around the model of well-being and interventions intended to improve the human condition and distill positive human functioning.

Keywords Topic modeling · BERTopic · Mental health · Reddit · LIME · Multiclass classification

✉ Amit Pundir
amitpundir@mac.du.ac.in

Pahalage Dona Thushari
thusharipahalage@gmail.com

Nitisha Aggarwal
nitisha@south.du.ac.in

Vajratiya Vajrobol
tiya101@south.du.ac.in

Geetika Jain Saxena
gsaxena@mac.du.ac.in

Sanjeev Singh
sanjeev@south.du.ac.in

[1] Department of Software Engineering, Delhi Technological University, New Delhi, Delhi 110042, India

[2] Institute of Informatics and Communication, University of Delhi, New Delhi, Delhi 110021, India

[3] Department of Electronics, Maharaja Agrasen College, University of Delhi, New Delhi, Delhi 110096, India

1 Introduction

The condition of one's mental health significantly impacts the person suffering and the community (Benrouba and Boudour 2023; Ji et al. 2022). According to the United Nations report, in 2022, one in eight people globally suffers from a mental health disorder (World mental health report: Transforming mental health for all - executive summary 2022a). The report has underscored the disparities in mental healthcare, underscoring the global prevalence of high mental health needs alongside frequently inadequate and insufficient responses. Furthermore, it has proposed measures to enhance mental healthcare systems, rendering them more accessible and effective for everyone. The delayed diagnosis of mental illnesses, resulting in delayed preventive interventions, has contributed to approximately 703,000 annual suicide-related deaths across various age groups, genders, and geographical regions. Presently, suicide stands as a leading

cause of mortality, surpassing other factors like malaria, AIDS, breast cancer, war, and homicide on a global scale. To identify high-risk demographics, nations should gather and scrutinize disaggregated data encompassing variables such as gender, age, and suicide methods. These data hold paramount importance in grasping the scope of the issue and facilitate the tailoring of interventions to suit the unique requirements of at-risk populations while adapting to evolving trends (Suicide data: Mental Health and Substance Use 2021). The diagnostic and statistical manual of mental disorders (DSM-V), published by the American Psychiatric Association, defines a mental disorder as a set of symptoms that cause significant disruption in a person's thinking, emotional regulation, or mental functioning that is consistent with psychological, biological, or developmental processes. Mental illnesses are typically accompanied by severe suffering or impairment in vital social, occupational, or other tasks (Stein et al. 2021). Unfortunately, more than 70% of individuals with mental disorders worldwide lack primary care and treatment due to limited resource environments (Kilbourne et al. 2018).

Mental health professionals study feelings, thoughts, behavior patterns, and other tests to diagnose illness. The availability of such resources is a critical constraint for conducting these studies, as health records of patients' emotional and intimate feelings are not shared due to privacy and confidentiality concerns. Data scarcity, therefore, restricts the research in this domain and severely limits designing accurate methods and techniques for diagnosing mental illness (Hanna et al. 2018). Moreover, most mental disorders have similar characteristics, making distinguishing between mental health diseases for diagnostic purposes difficult. In addition, the availability of mental health experts per million population, particularly in Low- and middle-income countries (LMICs), is deficient (Wainberg et al. 2017). Handling of few mental illnesses, such as depression, anxiety, and bipolar disorder, is challenging as they do not have cures. However, behavior therapy and medication can significantly control the illness and the quality of patients' lives may significantly improve. Therefore, reducing stress in peoples' lives is crucial by knowing the cause of illness and minimizing its impact on their daily lives.

Social network platforms allow individuals apprehensive about face-to-face interactions to share their thoughts and feelings with a larger community. These social media platforms are easy to access and do not discriminate based on age, gender, socio-economic status, race, or ethnicity. Communicating with a larger community of people suffering from similar concerns in online forums helps them regulate their emotions. Their interactions, in the form of texts, become an invaluable resource facilitating analysis of textual signs of psychological health problems (Dao et al. 2015; Kamarudin et al. 2021). Among the popular social media networks, the Reddit platform is preferred by young individuals for sharing their emotions as it provides anonymity and a handy platform for discussions on mental health issues (Boettcher 2021). Researchers using such data can analyze the emotions expressed to understand the mental issues the users suffer and design tools to find the types, stages and cost-effective solutions (Ren et al. 2021).

Recent years witnessed an abrupt growth in the analysis of social media data to investigate a variety of health issues ranging from the effects of allergies and Covid-19 (Huang et al. 2022; Zhou et al. 2021; Kathy et al. 2015) and sentimental analysis on Covid-19 vaccine (Alotaibi et al. 2023; Verma et al. 2023) to emotions and mental health conditions (Saha et al. 2016; Kim et al. 2020; Lin et al. 2023). Sentiment analysis is applied to identify the underlying sentiment or emotion expressed in a piece of text. The sentiment of a person's post, certain language and linguistic patterns can provide insights into their mental state. Natural language processing (NLP) analyzes such data and allows for diagnosing disorders and developing treatment strategies. Numerous methods have been developed to map the semantic relations between textual data (Qi and Shabrina 2023; Rizvi et al. 2023). Deep learning (DL) models recurrent neural network (RNN), long short-term memory (LSTM), transformer, convolutional neural networks (CNN), bidirectional encoder representations from transformers (BERT) and other hierarchical attention networks (HAN) have been applied to analyze text at different levels, such as document level and sentence level, for identifying mental health-related concerns. BERT and other transformer-based models employ attention-based mechanisms; hence, they offer several benefits over traditional models like CNN and LSTM in the context of NLP. This attention mechanism helps identify significant words, phrases, or patterns that contribute to the overall meaning or sentiment as they focus on different parts of the input text. They also consider the neighboring words and their relationships to allow for a better understanding of the meaning and semantics of a word within a sentence to capture contextual information effectively. Traditional models like CNN and LSTM process data sequentially and hence are limited in capturing long-range dependencies in text, while transformer models can capture dependencies between words regardless of their positional distance through parallel processing. Additionally, contextual methods, like the MentalBERT model, consider the context and relationships within the text to understand its meaning. They capture nuances and dependencies, making them suitable for mental health discussions. Non-contextual methods like Bag of Words and TF-IDF analyze words in isolation without considering the context. They extract features using statistical or rule-based techniques, providing insights without contextual nuances. Non-contextual methods may not capture the subtleties and dependencies present in the text as effectively as

contextual methods. However, they can still provide valuable insights, especially when the specific context is not crucial for the analysis. This paper uses recent developments in NLP to find the common underlying topics in communities with mental health issues using their social media interactions. Our work has three focuses of interest:

1. The extraction and evaluation of significant themes and subjects for sensitive mental health issues using topic modeling;
2. Comparative analysis of context-based and non-context-based machine learning (ML) classifiers to automate the classification process of textual data on mental health issues, and
3. Using model interpretability, analyze the secular associations present in the distinctive attributes.

Identifying the variables contributing to mental health issues based on social media interactions can reveal common factors. These factors may subsequently lead to determining new perspectives of interventions and strategies to achieve better mental health solutions. This research uncovered the main themes and patterns within mental health-related content derived from a social media platform, employing the topic modeling technique. Through these identified themes, we gain access to the underlying semantic structure embedded within the text. For instance, recurrent themes in the study encompassed topics like exams, school, college, friends, and relationships, highlighting their interconnectedness with mental health. Moreover, the outcomes revealed that context-driven models such as BERT and MentalBERT exhibited superior post-classification accuracy compared to traditional models like logistic regression (LR) and random forest (RF). To gain insights into the classification model results, we employed LIME, which indicated that BERT and MentalBERT classified posts based on contextual factors. In addition to this, our findings indicated that individuals with mental health concerns may potentially exhibit a likelihood of multiple mental disorders, a facet elucidated through the predictive capability of LIME explanations. The rest of the paper has been arranged as follows. Section 2 reports the related work, while the methodology is shown in Sect. 3. Section is about the findings of the study. Section 4 is the discussion related to the results obtained and their significance. Finally, in Sect. 5, the conclusions and future work are discussed.

2 Related work

Recent advancements in NLP and deep learning have positively influenced the analysis of social networking media interactions of communities with mental health issues. Social media data on platforms such as Twitter and Reddit (Liu et al. 2022) are of prime interest to researchers as they are real-time, readily available and help to reach a wider audience at a low cost. The research by Ji et al. (2022) describes a relation network with an attention mechanism to identify mental disorders and suicidal ideation with associated risk indicators. The SWMH (SuicideWatch and Mental Health) real-world dataset contains subreddit data divided into multiple classes, which we used in this study. They enhanced text representation and measured sentiment score and latent topics by lexicons. Their best-performing model for SWMH data achieved 64.74% accuracy for relation networks. For the same dataset, another study introduced pre-trained language models, namely MentalBERT and MentalRoBERTa (Ji et al. 2021) and achieved 72.16% best F1-score for the MentalRoBERTa model. These models were trained on domain-specific language for mental healthcare and are publicly available.

The work by Guntuku et al. (2017) investigated potential ways to use screening surveys on social media to predict mental health disorders. They detected symptoms associated with mental illness from Twitter, Facebook, and web forums and suggested that AI-driven methods can detect mental illnesses. An earlier study by Gemmell et al. (2019) investigated how to automatically recognize informal patterns in the language retrieved from online forums for borderline personality disorder patients as well as bipolar disorder patients. The top 10 phrases and terms were found to best describe each cluster using k-means clustering on the Reddit data. Another study by Kotenko et al. (2021) on the evaluation of the Mental Health of Social Network Users (Pushshift Reddit Dataset) calculated the emotion lexicon, used Latent Dirichlet Allocation (LDA) for topic modeling and classification using the fastText classifier and achieved 96% F1-score. Traditional approaches, such as LDA (Blei et al. 2003), failed to capture the semantic relationships and were not domain-specific, providing subjects irrelevant to the context.

Researchers widely prefer DL techniques in analyzing social media data to detect mental disorders. A study by Gkotsis et al. (2017) classified mental health conditions using feedforward and convolutional neural networks (CNN) based on the Reddit dataset. This Reddit dataset is further grouped through a semi-supervised technique to create subreddits of 11 mental health-related themes. The best-performing CNN model showed 71.37% accuracy in their study. Another study by Islam et al. (2018) on Facebook data detected depression using psycholinguistic measurements and ML algorithms, including the decision tree (DT), which outperformed all other techniques. Zanwar et al. (2022b) conducted a multiclass classification using hybrid and ensemble transformer models on the self-reported mental health diagnoses (SMHD) dataset and the Dreaddit dataset

using BERT, RoBERTa, and bidirectional long short-term memory (BiLSTM), achieving a macro F1-score of 31.40% across 5 folds.

Interpretability of models becomes crucial as classification alone is insufficient for sensitive applications such as mental health analysis. Explainable AI has emerged as an effective technique to address this problem (Gunning et al. 2019). One such technique, the local interpretable model-agnostic explanations (LIME) (Ribeiro et al. 2016), provides a factual implementation to explain respective predictions. This paper applied the LIME algorithm to present a trustable explanation. Hu et al. conducted a multiclass classification on Covid-19 related mental health data using the post hoc system and ante-hoc method to analyze and explain the factors that impact mental health during the pandemic (Hu and Sokolova 2021). Another recent study (Saxena et al. 2022) on CAMS (Garg et al. 2022) dataset used LIME and integrated gradient (IG) methods to find explanations for reasons related to inconsistency in the accuracy of multiclass classification. Our study found meaningful topics within mental health discussions to gain insights and interpret the findings of individuals suffering from mental health illness using a novel technique, BERTopic (Grootendorst 2022c). We analyzed the possible occurrences of topics for four clinically identified mental health issues, including bipolar disorder, anxiety, suicidal thoughts, and depression. The study has been extended using classification techniques to classify the text corpus into each category. Post classification, the trustworthiness of each model's prediction ability was investigated using local explanations for a given instance using explainable AI.

3 Methodology

3.1 Dataset

Given its novelty and accessibility, we have examined the Reddit SWMH (Ji et al. 2022) dataset. The dataset comprised texts from mental health-related reddit subreddits of anxiety, depression, bipolar, and suicidewatch, a total of 54,412 texts in five classes. This dataset is anonymous; hence, the absence of any identifiable details regarding the individuals in the dataset ensures that the privacy and confidentiality of the participants remain uncompromised. The SWMH dataset's purpose is to identify associated mental health conditions-related variables. This study removed the 'self.offmychest' class and duplicate texts. This class was removed as it did not indicate any direct relation to mental health illnesses, such as depression, anxiety, bipolar disorder, and suicidal thoughts. Our study examined 46,103 instances from the four classes, namely self.Anxiety, self.bipolar, self.depression, self.SuicideWatch. This dataset was analyzed to find themes and patterns of suicidality and mental illnesses.

The total number of texts evaluated for each dataset is shown in Fig. 1. We assessed and contrasted the data available on social media for various mental disorders with recurring themes and patterns associated with mental illnesses. The text samples in Table 1 represent each category.

Word clouds were created to visualize the importance of words in the context. The word cloud generated and shown in Fig. 2, using the text corpus, reveals the frequently used words "feel," "life," and "one" to convey emotions. Words such as 'anxiety,' 'depression,' and 'people' tend to stand out since the users have been sharing texts with phrases like "*I feel anxious around people*" and "*Is there anyone to talk about depression?*" for example. Social media users frequently utilize social platforms allowing for more open discussions about delicate subjects like mental health (Yazdavar et al. 2018).

3.2 Process architecture

The process architecture of this experiment consists of preprocessing texts, topic modeling, classification and local explanations, as shown in Fig. 3. The pre-processing pipeline was further segmented into noise removal and normalization. As for noise removal, the datasets were processed to remove duplicate text, links, text in square brackets, terms with numbers, and lowercase text conversion. In addition, stop words and largely standard English terms lacking crucial information and meaningless and misspelled words were removed. The NLTK library (natural language toolkit) was initially used for this purpose, and the tokenization step was then completed. Tokenization is a crucial process of dividing a text into token-sized pieces allowing for interpreting the text's meaning via word analysis. Several library functions, such as sci-kit-learn countvectorizer for ML models and Tensorflow Tokenizer for CNN and LSTM models, were

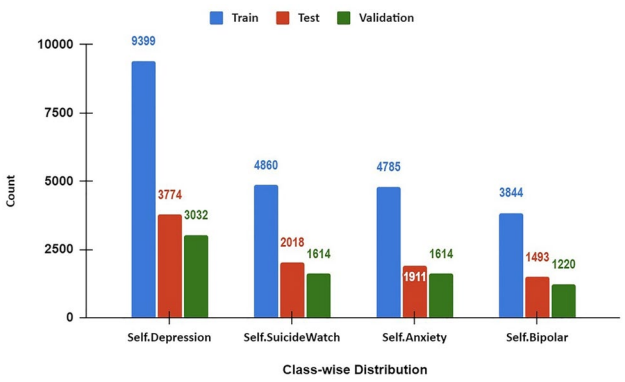


Fig. 1 Class-wise distribution of processed training, testing, and validation sets

Table 1 Samples of text from each class

Label	Text sample
Anxiety	Why do my arms, and muscles jerk when I feel anxious? When my anxiety is particularly bad, I find that my muscle control is horrible. When I try to grip something, my muscles will simply twitch. It feels and appears incredibly strange. Even turning my head in one direction causes it to jerk. I really hate this. It gives me the impression that there is a problem with me
Bipolar	Having recovered from a manic episode and considering scheduling a meeting? Since my fiancé broke up with me two months ago, I've apparently been experiencing manic episodes, and now I'm beginning to collapse
Depression	Right I must speak with someone. Every time I try to talk to my friends about my problems, they either tell me I'm acting silly or that things don't really matter, that nobody cares, or that I'm just being a little bitch. They could be correct, but those things have been bugging me for about two to three months now, and I really need to get it off my chest because it's getting exhausting. Although I frequently consider suicide, I lack the courage to carry it out
SuicideWatch	After a few days of abstinence, I cut today. Even though it wasn't much time, I felt pretty proud of myself for not doing it. And right now, slashing myself and taking my own life is all I can think about. I should have done it a week ago, but I decided not to, which was a terrible error. I overdosed four times in the last five months, and I wish one of them had killed me. I know if I leap off the nearby overpass of the highway, I'll get better results. This world is not suited for me. I have no place here. I just want to fall asleep and stay asleep forever

Fig. 2 Word cloud for training data to visualize important words in the context of mental health

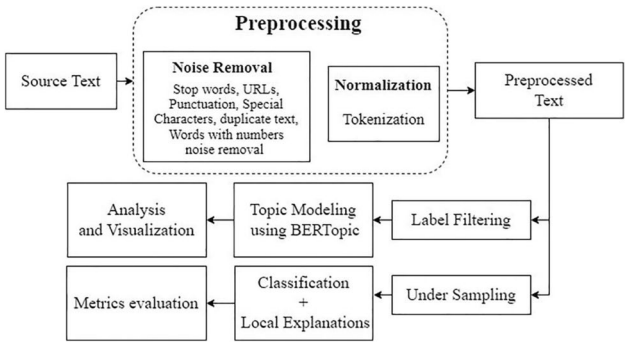


Fig. 3 Architecture of the approach used in this study to identify mental health concerns

used to complete this step. The Transformer models were used with their respective tokenizers provided by the HuggingFace platform. Normalization, a pre-processing step used to enhance the text's quality and prepare it for machine processing, was done. While conducting the classification, undersampling and class weights were applied to deal with the class imbalance.

3.2.1 Topic modeling

By examining the terms in the source texts, topic modeling seeks to identify the themes that permeate a corpus. Topic modeling is an unsupervised ML method of identifying

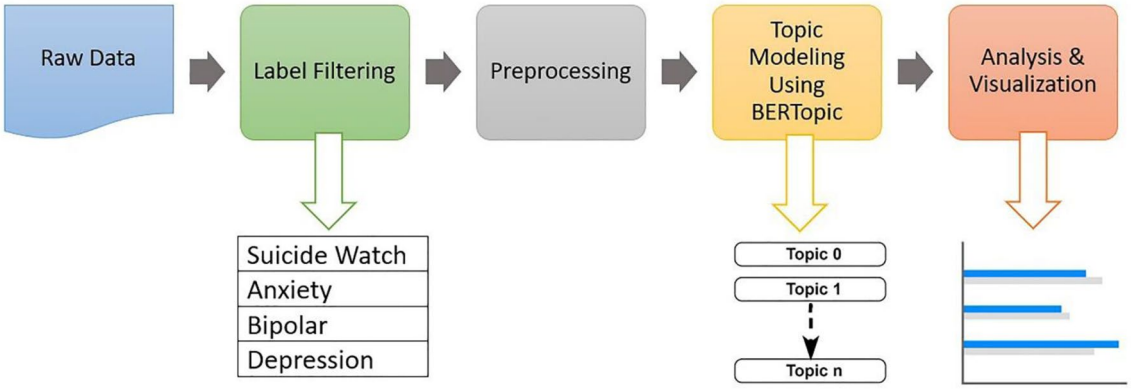


Fig. 4 Architecture of the approach used in this study for topic modeling

or extracting topics by spotting patterns, much like clustering algorithms that separate data into various sections. Transformers, C-TF-IDF and BERTopic, use dense clusters to generate simple to understand topics while preserving critical terms from the topic descriptions. Earlier BERTopic-based studies (Sangaraju et al. 2022; Abuzayed and Al-Khalifa 2021) found that it is adaptable and offers distinct topics compared to LDA. BERTopic has expanded the classical cluster embedding approach by utilizing cutting-edge language models and a class-based TF-IDF(C-TF-IDF) process to create topic representations. The model uses three phases to generate topic representations. Each document is first transformed using a trained language model into its embedding representation. The dimensionality of the generated embeddings is then decreased before clustering them to improve the clustering procedure. Due to the stochastic nature of the UMAP method used for dimensionality reduction while clustering, the model gives different results for training instances. Finally, topic representations are retrieved from the document clusters using a customized class-based form of TF-IDF. The model is significantly more flexible and easier to use because the grouping of documents and generating topic representations is carried out independently (Grootendorst 2022c). C-TF-IDF is utilized in the third stage to extract essential terms at each cluster on a class-based term frequency or inverse document frequency. In BERTopic, TF-IDF was modified to operate on a cluster/category/topic level rather than at the document level to accurately represent the subjects from the bag-of-words matrix. This modified TF-IDF representation, known as C-TF-IDF, considers the differences between documents in each cluster by treating each cluster as a single document rather than a collection of documents. The frequency of the word x in class c, where c is the cluster that we previously established, is extracted. The class-based TF representation is the outcome of this. L1-normalization is used in this representation to consider the variations in topics. A customized version of the algorithm rather than the standard TF-IDF approach is utilized in BERTopic, allowing for a far better representation. The modified algorithm Grootendorst (2022c) proposed for this computation is shown below.

$$W_{\text{tc}} = \text{tf}_{t,c} \cdot log\left(1 + \frac{A}{\text{tf}_t}\right) \tag{1}$$

Here, the term frequency represents the frequency of term t in class c in Eq. (1). The group of documents combined into a single document for each cluster is class C in this instance. Then, evaluating how much knowledge a term contributes to a class, the inverse document frequency is substituted with the inverse class frequency. It is the frequency of term t across all classes divided by the logarithm of the average number of words per class A. It is beneficial when dealing with extensive collections of documents where it is crucial to identify the most important terms for each class.

In topic modeling, a topic is a cluster of words. These words are selected based on their statistical significance in the model. The importance of words is determined by their frequency of occurrence in the cluster and their co-occurrence with other words in the same context. Therefore, it is very helpful for interpreting the groupings produced by any unsupervised clustering method. Bar charts of the C-TF-IDF scores of each topic provide two insights: Scores for each word in each topic's C-TF-IDF and a comparative study of each topic's distribution. The process used for Topic Modeling is depicted in Fig. 4.

3.2.2 Explainable AI

An emerging subset of AI called explainable AI (XAI) focuses on the readability of ML models. An explainable AI model is a set of steps and strategies that enables one to comprehend and accept ML algorithms' results. It entails specifying the model's correctness and transparency and the outcomes of decision-making assisted by the model. LIME (Ribeiro et al. 2016) is a model-independent interpretability method for individual local predictions. Model agnostic refers to the ability to generate explanations for any DL or ML model. The degree to which a person can decipher the reasoning behind a black box model's choice is known as interpretability. Instead of creating an explanation for the whole, LIME's general operating premise is to localize the problem and explain it.

3.3 Evaluation metrics

The evaluation metrics for classification models are accuracy, precision, recall, and F-score. Accuracy is the ratio of the total number of positives to the total number of classes. Precision is the ratio of true positives to the total number of predicted positives. The recall is the ratio of true positives to the actual number of positives. We can get a harmonic mean of these measures using the F1-Score that considers precision and recall. When there is a significant class imbalance, this is very helpful. The formulas are as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{2}$$

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

where TN, TP, FN, and FP represent true negative, true positive, false negative, and false positive, respectively in Eq. (2–5).

In general, classification models are assessed using the above metrics, but since data that have been collected and annotated may have bias compared to the real world. Additionally, LIME is employed to explain by enhancing interpretability. In our case, the SWMH data set is based on the mental health-related subreddits where nuances of language and the usage of the words need a deep understanding of the context in which they were used. Therefore, the traditional metrics might not accurately reflect the primary goal of developing the text classification model. In addition to such metrics, analyzing individual predictions using LIME can be helpful. However, there is no formal agreement on what interpretability signifies in ML and its measurement methods (Molnar n.d.). A local prediction is defined as $L(f, g, x)$ applying the regularization parameters $\Omega(g)$. The L indicates the minimized square loss function and g represents the model applied to class G (class G is defined as the set of models that have the capacity for interpretability). The faithfulness of the explanation g to the original Explanation(x) is measured to get an explanation of a local point x. The formula is as follows:

$$\text{Explanation}(x) = \arg\min_{g \in G} L(f, g, x) + \Omega(g) \tag{6}$$

where $\arg\min_g$ is defined as the value of argument 'g' for which function $GL(f, g, x)$ attains its minimum.

3.4 Model definitions

Multiclass classification, often known as multinomial classification, categorizes cases into three or more classes in statistical classification and ML. Each data sample can be put into a specific class. A data sample, however, cannot concurrently be a member of more than one class. In other words, the classes in multiclass problems are mutually exclusive.

3.4.1 Logistic regression (LR)

By default, logistic regression can only be used to solve binary classification issues. As a result, it has to be modified to enable multiclass categorization issues. Studies (Pranckevičius and Marcinkevičius 2017) have shown that Logistic regression can be used to achieve higher performance in multiclass text classification. Logistic

Regression is helpful in this situation as it uses a sigmoid function to output a probability between zero and one. In this study, we have adapted a logistic regression model to recognize and forecast a multinomial probability distribution known as multinomial logistic regression.

3.4.2 Random forest (RF)

As an extension of bagging, the random forest (Breiman 2001) algorithm randomly chooses portions of the features utilized in each data instance. Decision trees, which are the foundation of the random forest model, are susceptible to class imbalance. Every tree is supported by a "bag," representing a uniform random data sampling. As a result, a class imbalance will, on average, bias each tree in the same direction and magnitude.

3.4.3 Stochastic gradient descent (SGD)

SGD is an optimization approach for updating a model's weights while being trained. The algorithm updates the weights based on the difference between each input text's actual and projected class. By merging various binary classifiers in an "OVA" (one versus all) framework, it can be used for multiclass classification. A binary classifier is learned for each class that can distinguish it from all other classes. For large datasets, SGD provides a quick and effective optimization approach.

3.4.4 Naive Bayes (NB)

Based on the Bayes theorem, the likelihood of a class given a set of features is proportional to the likelihood of the features given the class. Naive Bayes models are frequently employed for text classification problems since they are quick and simple to implement.

3.4.5 XGBoost (XGB)

GradientBoost was first introduced by Chen in 2016 and XGB (Chen and Guestrin 2016) is its improved version. A gradient boosting framework is used by the decision tree-based ensemble ML algorithm XGB to handle missing values in the training phase and control overfitting and split finding.

3.4.6 Long short-term memory (LSTM)

LSTM is a form of recurrent neural network (RNN). For text data sequences where the word order matters, LSTM models are very well suited. Long-term dependencies can be captured by LSTMs, which also have the ability to handle sequential dependencies in the data.

3.4.7 Convolutional neural network (CNN)

Convolutional Neural Networks can be utilized to solve text classification issues in addition to picture and video analysis. When classifying text, the input is handled like a picture, with each word acting as a feature. The fully connected layers are utilized for prediction, whereas the convolutional layers extract features from the input.

3.4.8 BERT

BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al. 2018) is a pre-trained DL model. BERT models are customized for certain classification tasks after being pre-trained on a sizable corpus of text data. BERT models are ideally suited for text categorization issues since they have a reputation for capturing the context of the words in a text.

3.4.9 MentalBERT

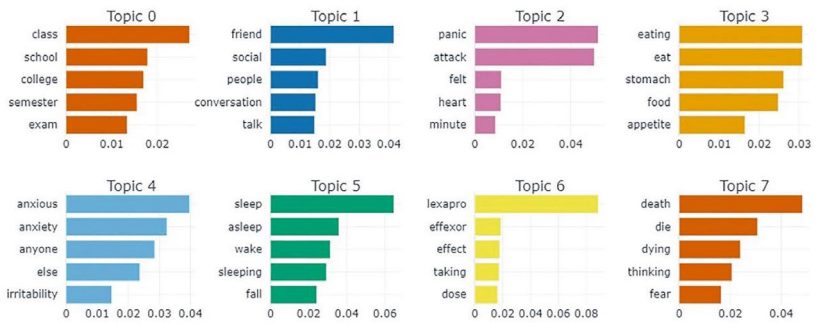
MentalBERT (Ji et al. 2021) uses the BERT model's pre-trained knowledge for contextualized language representations related to mental health. The classification head is trained using a sizable corpus of annotated text to determine the class of a given text document. The MentalBERT model's weights are adjusted during fine-tuning for optimum text categorization performance.

4 Results and discussion

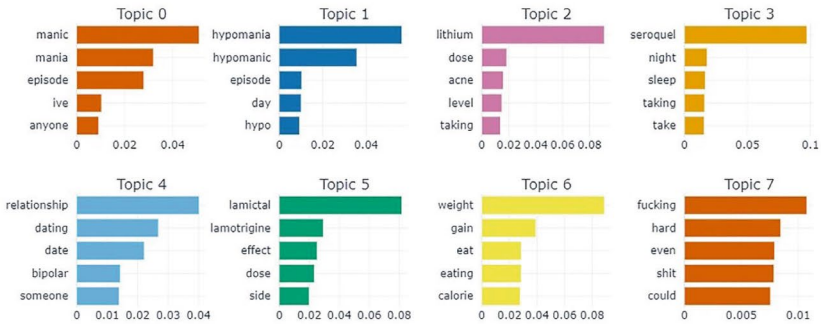
4.1 Topic modeling

We investigated four commonly identified mental health problems, examining the recurring themes and vocabulary that emerged. Plots for the eight most popular topics under each corresponding mental health condition are shown in columns of Fig. 5. Each topic cluster's five most common words are plotted on each bar chart. Each row of the bar graph with the C-TF-IDF score shows the most common terms from each cluster. It can be concluded from the results that the subjects like *class, school, holiday or exam* can trigger the symptoms of mental health problems. These results can be used to aid the therapy and medication processes and to empathize with mental health patients.

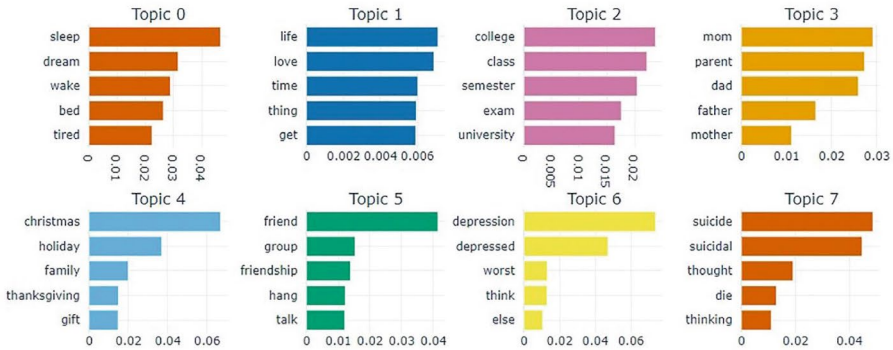
Recurring themes are observed in each class. For instance, topics related to college, work, and relationships frequently appear in every category examined. Demographically speaking, the reason for this can be that Reddit users are mainly young, tech-savvy people who use social media platforms more often to discuss their issues daily. Thus, this study can be extended to focus exclusively on the problems



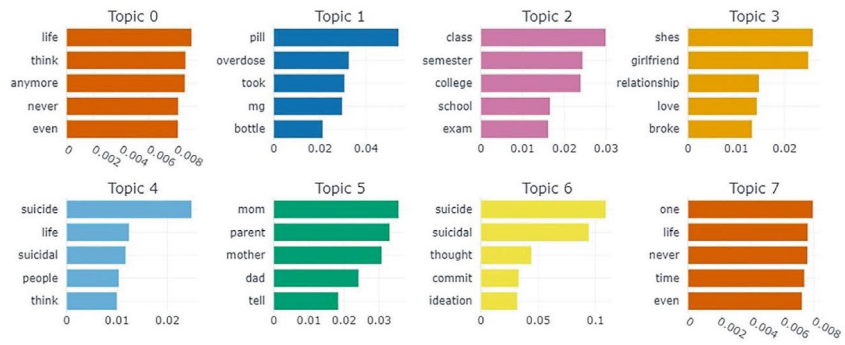
(a) Anxiety



(b) Bipolar



(c) Depression



(d) Suicide

Fig. 5 Graph of recurring words with their C-TF-IDF scores for each class

of youth in modern society that trigger mental health issues and trauma. The themes about bipolar disorder stand out from the rest, as people with bipolar disorder seek professional help compared to the rest. Thus, the topics related to medication and therapy were prominent in that category.

However, it is noteworthy that the BERTopic model has several shortcomings (Grootendorst 2022c). It presumes that each document contains one topic, while in reality, documents may have numerous topics. The word corpus was filtered using the label mental health issue, and five similar motifs in topics were found through qualitative analysis. In each health problem, the topic clusters were related. We could gather subjects that were self-contained and coherent as a result. Given the highly topical nature of Reddit's subreddits, it is expected that some topics, like bipolar disorder studies, are more specific to a particular subreddit than others. The present research facilitates comparing and examining overlapping vocabulary and underlying semantic attributes over time. The relative frequency variation of various clusters can be observed in a given mental health condition, such as depression or suicide. The procedure outlined also enables us to assess conversational shifts related to various topics within a related category and compare it with the rest. It would be helpful for mental health practitioners to understand how various conversations evolve or how they respond to particular events differently.

4.2 Classification

Classification techniques were used to perform Multiclass classification of the text corpus into each category. Since the data set is considered mutually exclusive, each text is only in one category. However, the cases may fall under two categories simultaneously. For instance, a depressed patient might also be suffering from suicidal tendencies at the same time. In addition, in some instances, it was noticed that texts must be of a substantial length to gain enough context to classify correctly.

Using several ML, DL and transformer models for classification allowed us to examine context-based against non-context-based classifier performance. The existence of an imbalanced class distribution within this dataset has repercussions on the model's effectiveness. Because the model exhibits a bias toward the majority class, its capacity to effectively learn from the minority class is compromised. To tackle this challenge, both undersampling and class weight techniques were applied to the dataset. Figure 6 showcases a comparison of the outcomes before and after addressing the data imbalance, providing a visual representation of the effects of these strategies. The domain-specific MentalBERT outperformed all other models (context and non-context-based methods). Table 2 displays the average performance evaluation of multiclass classifiers. The models performed better after undersampling the majority-class data, with the best-performing model MentalBERT achieving 76.70% mean accuracy (best-performing results are in bold in Table 2), outperforming the results published in previous studies (Ji et al. 2022, 2021) by more than 4%. To evaluate the model's classification performance, an accuracy error bar graph (Fig. 7) has been plotted for all nine classification models. Classical Models such as NB, RF, and transformer-based models BERT and MentalBERT have shown very low variance in accuracies compared to CNN and LSTM. This low variance shows that the model's performance is consistent and generalizes well across all the subsets of the dataset. Confusion matrix has also reported in Fig. 8 to understand the statistics of performance for best-performing model. MentalBert has a identify the true labels and false positives and negatives are very less for most of the classes. In literature (Ji et al. 2022), a relation network (RN) based on

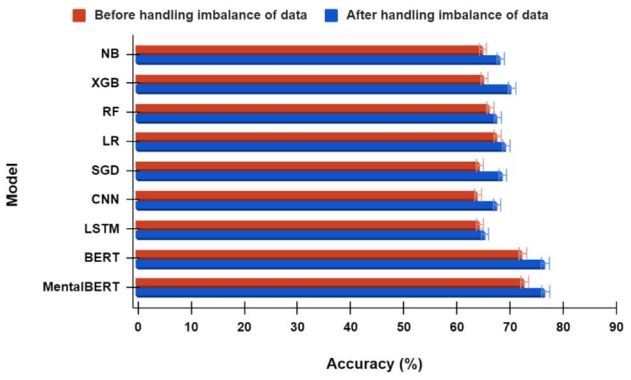


Fig. 6 Comparison of models on the basis of accuracy before and after handling of imbalanced data

Table 2 Results of classification performance metrics for all nine models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
NB	68.23	68.87	68.23	68.32
XGB	70.38	74.49	68.13	70.53
RF	67.69	67.99	67.99	67.40
LR	69.30	69.92	69.30	69.55
SGD	68.63	71.54	68.63	68.77
CNN	67.56	68.54	66.35	67.41
LSTM	65.29	67.40	61.40	64.20
BERT	76.62	76.52	76.56	76.56
MentalBERT	**76.70**	**76.66**	**76.69**	**76.63**

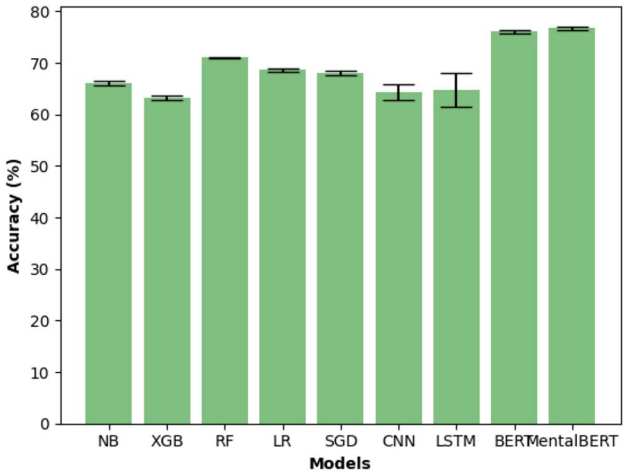


Fig. 7 Error bar graph for accuracy on all nine classification models

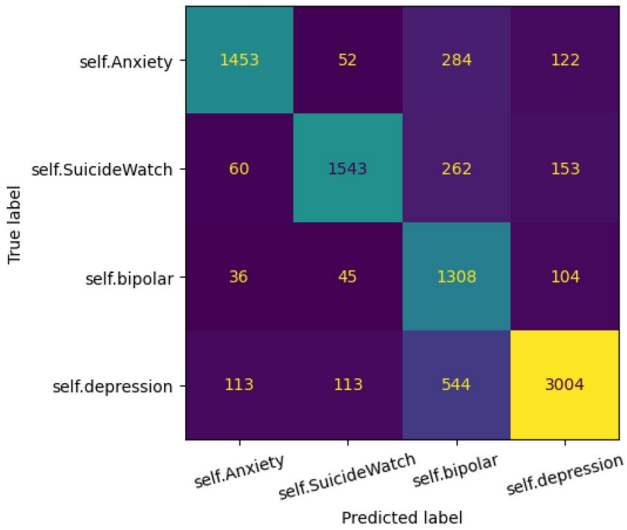


Fig. 8 Confusion matrix for best-performing model (MentalBERT)

the attention technique was used to classify the text. This RN model achieved 64.74% accuracy and 64.78% F1-score. Another study (Ji et al. 2021) applied MentalRoBERTa on the SWMH dataset and attained 72.16% F1-score.

One more experiment was designed to understand further the decision various models took. In this study, nine models are considered for classification. The SGD model outperformed four models (RF, LSTM, XGB and NB) but lagged behind the other four models (BERT, CNN, LR and MentalBERT) in terms of accuracy. Hence, 50 misclassified instances of SGD from all classes were selected and checked for their LIME explanations ("6."). Also, the other 8 models were applied to check their efficiency on these misclassified instances. As shown in Fig. 9, MentalBERT correctly classified twenty-four instances out of 50, whereas RF could correctly classify only sixteen. These posts have mixed topic themes and class-representative keywords that may confuse models, yet MentalBERT performed well compared to all other classification techniques used in this study. Despite having a smaller training corpus size compared to BERT, MentalBERT demonstrates comparable or superior performance on mental health datasets. This is due to MentalBERT's capability for capturing the specific context and sentiments prevalent in mental health-related text. While BERT demands substantial computational resources for training and inference due to its general-purpose applicability and extensive pre-training corpus, MentalBERT's specialized training results in reduced resource utilization during both the training and inference stages. This decreased resource overhead can prove advantageous in situations where computational resources are limited or when deploying the model on devices with constrained memory capacities.

An additional dataset ("depression" 2021) extracted from Facebook comments and posts on mental health-related issues, is analyzed within the same experimental setup to validate the proposed approach. This dataset comprises 6982 social media posts and is publicly available on the Kaggle platform. The dataset consists of two classes and has been annotated using a majority voting approach involving undergraduate students. All nine classification models were applied to this dataset, and the results are presented in Fig. 10. The classification outcomes were compared with the results reported in the literature (Hassan et al. 2021) and Kaggle notebooks to ensure consistency and reliability.

4.3 Explainable AI

LIME determines whether a system produces insightful justifications for its classification. The features employed in this work can be used to support ML assessments and contribute to the design of a reliable user interface. Figure 11 shows the

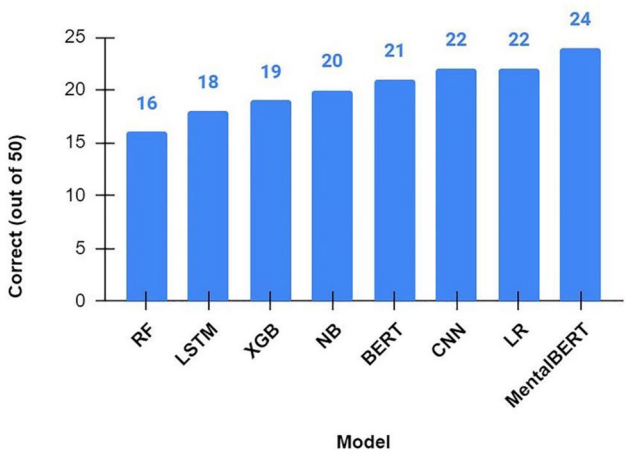


Fig. 9 Count of correct posts classified by all models except SGD

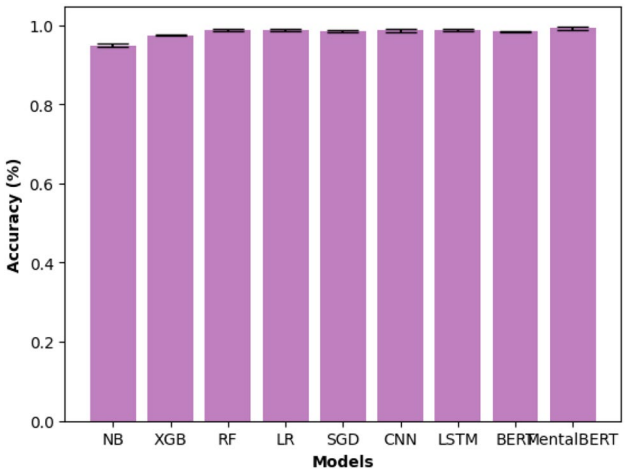


Fig. 10 Error bar graph for accuracy on all nine classification models on the Kaggle dataset

LIME algorithm outcomes for each model. The total class probabilities, the top features with their probabilities, and automatically highlighted features in the sample input text using LIME are considered to understand the results.

The class probabilities for a given model range from 0–0.99%. For instance, as shown in Fig. 11, a test text (I want to turn into an alcoholic and die in 6 years) taken for the depression class for the XGB model gives a 0.52% probability of '*self.suicidewatch*' label because it has the word *die* and the top features with their probabilities explain each class probability indicating the content is biased toward *suicidewatch* class. According to the classification from the XGB and NB models and explanations from LIME, the text consists of *suicidewatch* class (Fig. 11a, b). However, the true class (*self.depression*) of the text sample was not the same as the model's prediction because the sample text has few words (less information) and does not contain class-representative keywords (depression or depressed). The DL—LSTM model performed well for this sample text, while CNN could not because it gets confused with the words like '*deleted'* and '*die'*. RF, LR and SGD also performed well in this text but failed in others (see "6.").

Context-based models have tried to understand the context and give a higher probability to the true class (Fig. 11i). To our knowledge, the class-representative words or repetitive words (Fig. 5) have contributed to the classification of non-context-based models. While context-based models such as BERT and MentalBERT have classified text based on the overall context of the post, not on a few representative words.

The approach used in this study demonstrates a thorough and thoughtful process to enhance the performance and interpretability of NLP models. Firstly, data pre-processing techniques were implemented to clean the input data and reduce noise by removing stopwords, URLs, and punctuation. Additionally, duplicate texts were eliminated to ensure that the model was not biased by redundant information. Data imbalance, which could have led to biased predictions, was addressed through data undersampling and class weight assignment. Undersampling techniques helped balance the distribution of different classes, while class weights ensured that the model received adequate exposure to the minority class. It significantly enhanced the model's capacity to accurately predict all classes, regardless of their prevalence, compared to the previous study (Ji et al. 2021). This approach was further enriched by incorporating BERT for topic modeling. BERT, as a context-aware model, extracted meaningful themes and intricate details from text data related to mental disorders. By harnessing the capabilities of these models, we gained valuable insights into the underlying patterns within the data, thereby advancing our comprehension of mental disorders and their subtleties. Employing LIME for an explainable NLP analysis bolstered the transparency and reliability of this approach. LIME enabled us to analyze and elucidate the model's predictions by pinpointing crucial features and elucidating their contributions. It facilitated a deeper understanding of the model's decision-making process, rendering it more understandable and trustworthy in contrast to the prior study (Ji et al. 2021). Our approach amalgamates data pre-processing, the rectification of class imbalance, the utilization of BERT for topic modeling, the application of LIME for explainable NLP, culminating in a comprehensive workflow that not only enhances the model's predictive performance but also furnishes transparency in its results.

5 Conclusion and future work

This work captured complex relationships in a wide range of textual data, notably social media posts, using NLP techniques. It examined the distribution of topics in the Reddit-based online community for mental health. The social media data was analyzed for the significant themes related to mental health issues using BERTopic. The topics like relationships, exams, and school were found to impact mental health conditions directly. In addition, the themes that emerged from each category's topics were found to fit into several identifiable patterns. A few patterns related to school, friends and exams are recurring, frequently employing vocabulary from dialogues about mental health. The differences and similarities among the themes covered by the corpus of texts in each community were examined.

A comparative analysis using nine classical state-of-the-art classification techniques was done to classify the mental health disease that may help professional mental health therapists by automating the textual analysis

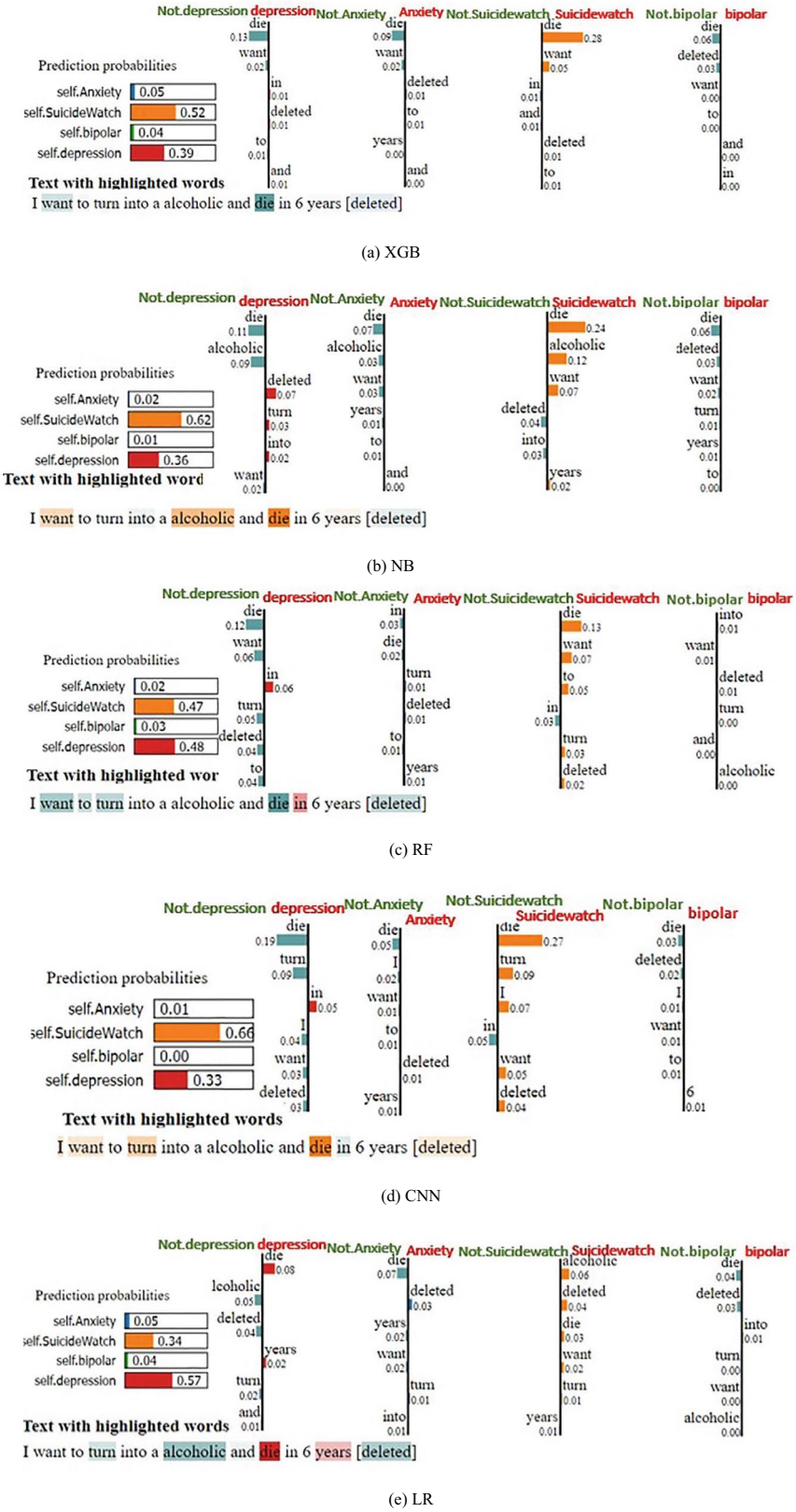
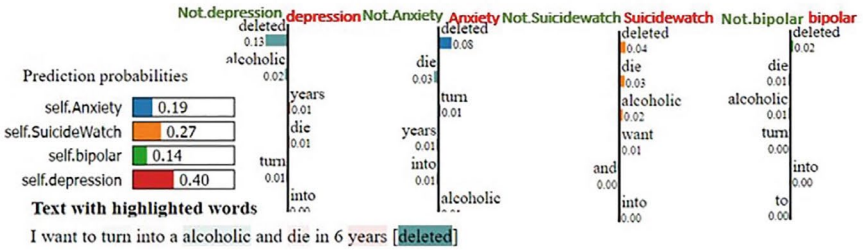
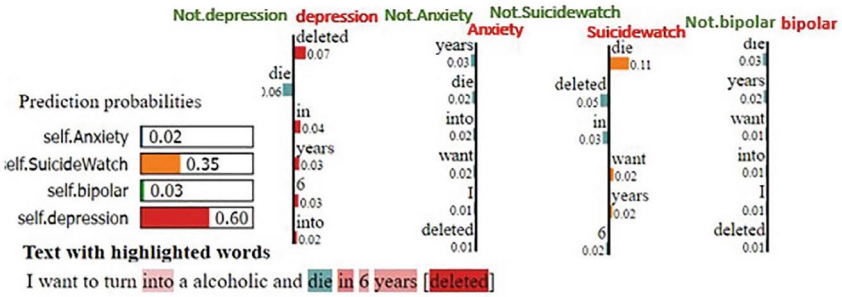


Fig. 11 Prediction probabilities of all nine models using LIME technique to explain model's decision

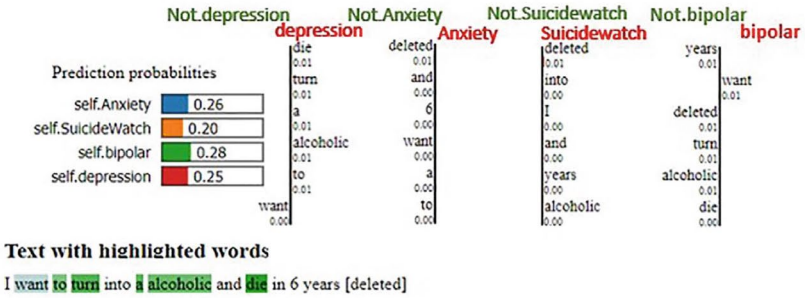
Fig. 11 (continued)



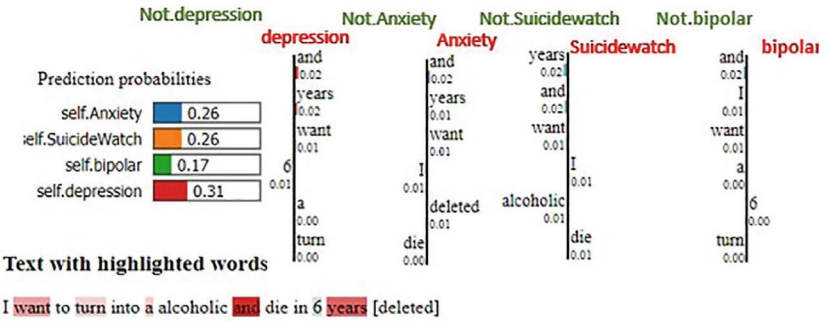
(f) SGD



(g) LSTM



(h) BERT



(i) MentalBERT

process. The transformer model, MentalBERT performed the best, considering the significance of context, and achieved 76.70% accuracy, significantly higher than the reported accuracies. The performance of the models and techniques was evaluated using the explainable AI technique–LIME. It aids in clarifying why a model is producing particular predictions and fostering trust. The results from this study can help identify users who may be at threat of mental health disease. This approach can differentiate users with potential mental health diseases, namely bipolar, depression, anxiety, and even suicidal ideation, based on symptoms extracted from their posts. The recognition of these recurring patterns and themes holds significant potential to assist mental health professionals in their practice. By identifying language patterns that signal the onset of mental health challenges, professionals can intervene proactively and extend support to individuals facing such risks. This early intervention has the capacity to curtail the progression of these mental challenges into more severe conditions. Mental health professionals can apply these insights to fine-tune their interventions and outreach efforts, increasing their effectiveness. Moreover, this study can serve as a valuable resource for policymakers, furnishing crucial data to shape mental health policies and initiatives. Comprehending the central themes and issues impacting mental health, as elucidated in this study, can inform the development of targeted mental health programs. In this research, we utilized solely English-language Reddit posts as our primary social media data source, which represents a constraint in our study. Enhancing the findings could involve incorporating information from a broader array of social media platforms beyond those using English, as well as interactions with individuals from a more extensive range of socio-economic backgrounds. This approach can be applied to larger datasets and more variable investigations in the future to improve the generalization of the study. It can be developed to explore the relationships between the uncovered topics related to mental health, like potential suicide identification or drug addiction.

Appendix 1

A sample text (lets called 'A') '*Feelings of excitement when thinking about suicide The past few days I've been dwelling on stupid past mistakes and getting a lot of invasive suicidal thoughts, but lately when I get them I start to feel excited like I'm going some place like an amusement park or something. When I think about killing myself it feels so right, rather than feeling dark and wrong. My heart starts racing really fast and I get tunnel vision at times and think about hanging myself, but then I immediately calm down and the thoughts stop after just 10 s or so. I'm not sure if this is common or not.*' is taken from test set belongs to *self.depression* class has classified for all nine classification models. In Table 3, classification insights have been reported for all the models used in this study. XGB, NB, CNN, SGD and LSTM wrongly classified the sample text as self.SuicideWatch (because it contains keyword *suicide*), self.Anxity(it contains words *racing* and *calm*), self.bipolar, self.SuicideWatch, and self.Anxity, respectively. However, RF, LR, BERT, MentalBERT classified the this text correctly as *self.depression*.

The true class for another sample text (lets called 'B') '*Back Well, after my close call that was the subject of my first post, here, I vowed to do something with my life, but it looks like I've gone around in a big fucking circle. I've liked her since high school, I've spent years becoming best friends with her and her boyfriend and she just doesn't fuckin want it. There is literally no end to this Romeo style Petrarchan lover bullshit but as much as I'm self-aware and introspective, I'm also a slave to my biochemical pitfalls. I'm seriously considering it.'* is self.SuicideWatch. Non-context-based models XGB and LR correctly identified this text where as deep learning models CNN and LSTM both misclassified sample text 'B.' Both context-based models BERT and MentalBERT also predicted actual class as shown in Table 4. Although BERT and MentalBERT gave same prediction probabilities for all classes but the selected different words to predict true class.

One more sample text (lets called 'C') from class self.depression has been taken '*depression and anxiety Attack!! Earlier I was feeling depressed and now I am having a anxiety attack. my stomach is churning, and I just want to lay down and die in all Honesty I feel like bashing my head against a wall maybe then I wouldn't feel like this. I have a sense of dread. I don't have any clue why I am feeling this way, but I just want to cry myself to sleep at least. but no. that's not even a possibility because I can't stop fidgeting and stirring and my mind won't calm down. not even breathing in and out helps.'* to understand the model's classification decision. The text contains both anxiety and

Table 3 Lime explanation for sample text A of each model

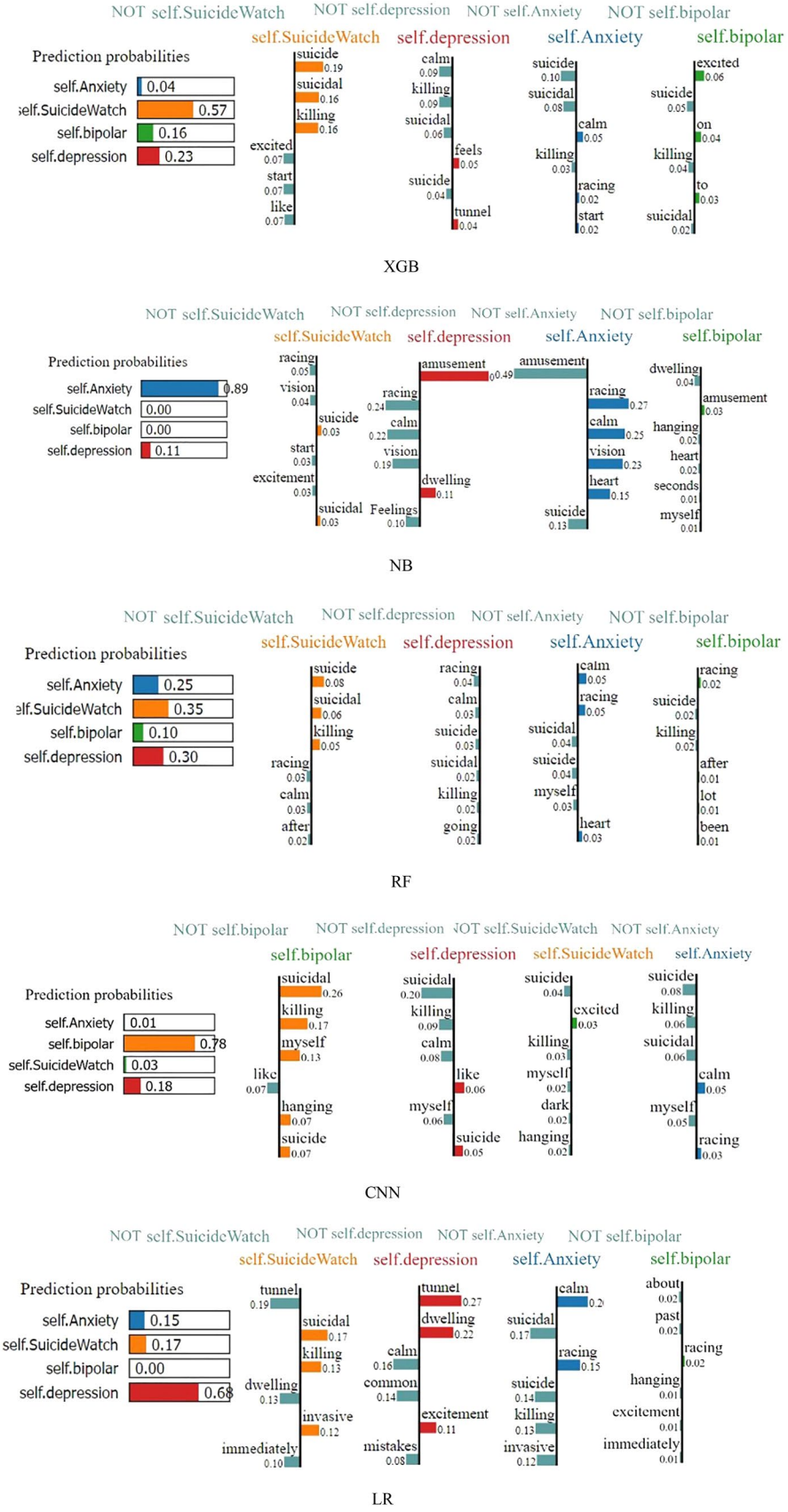


Table 3 (continued)

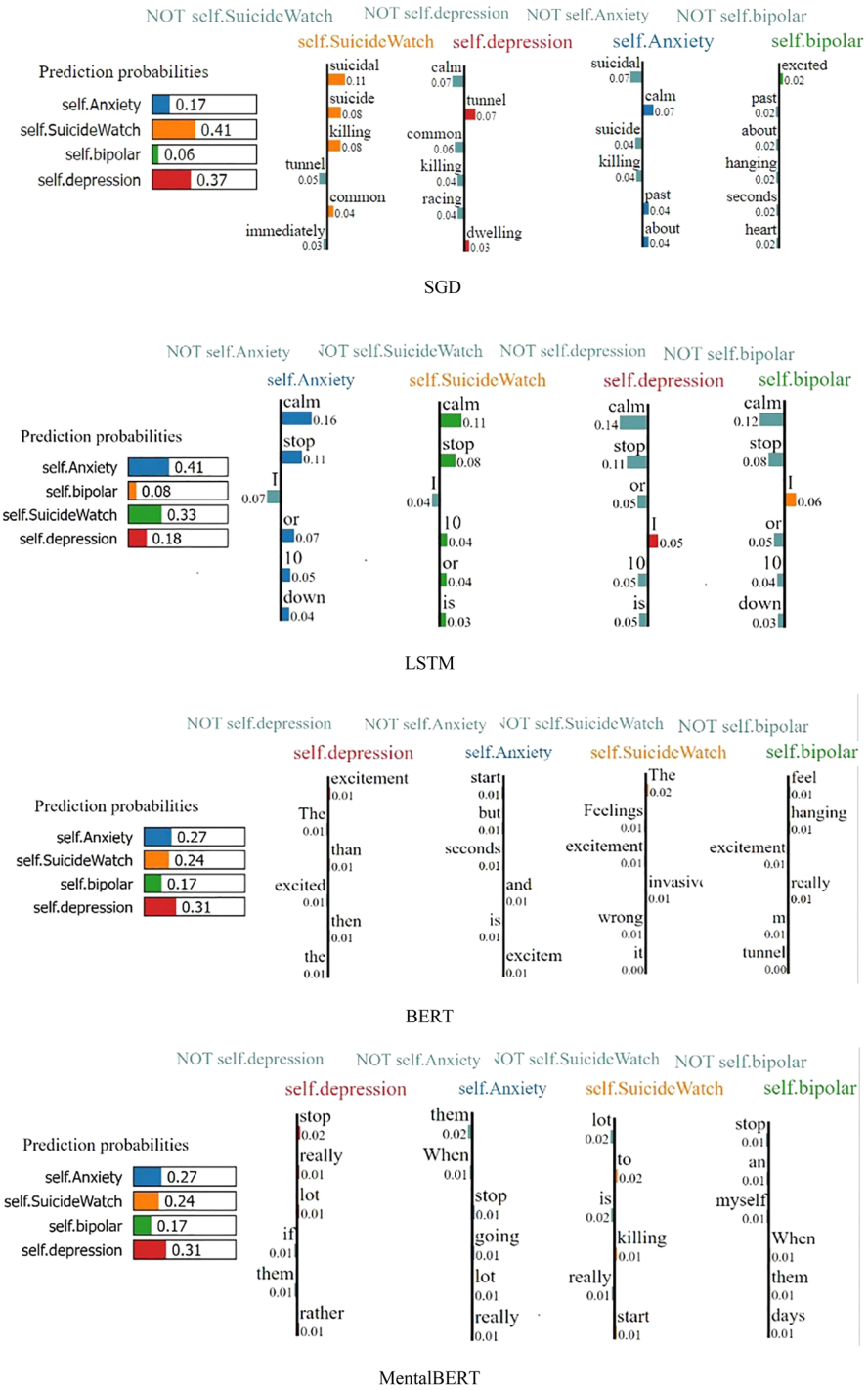


Table 4 Lime explanation for sample text B of each model

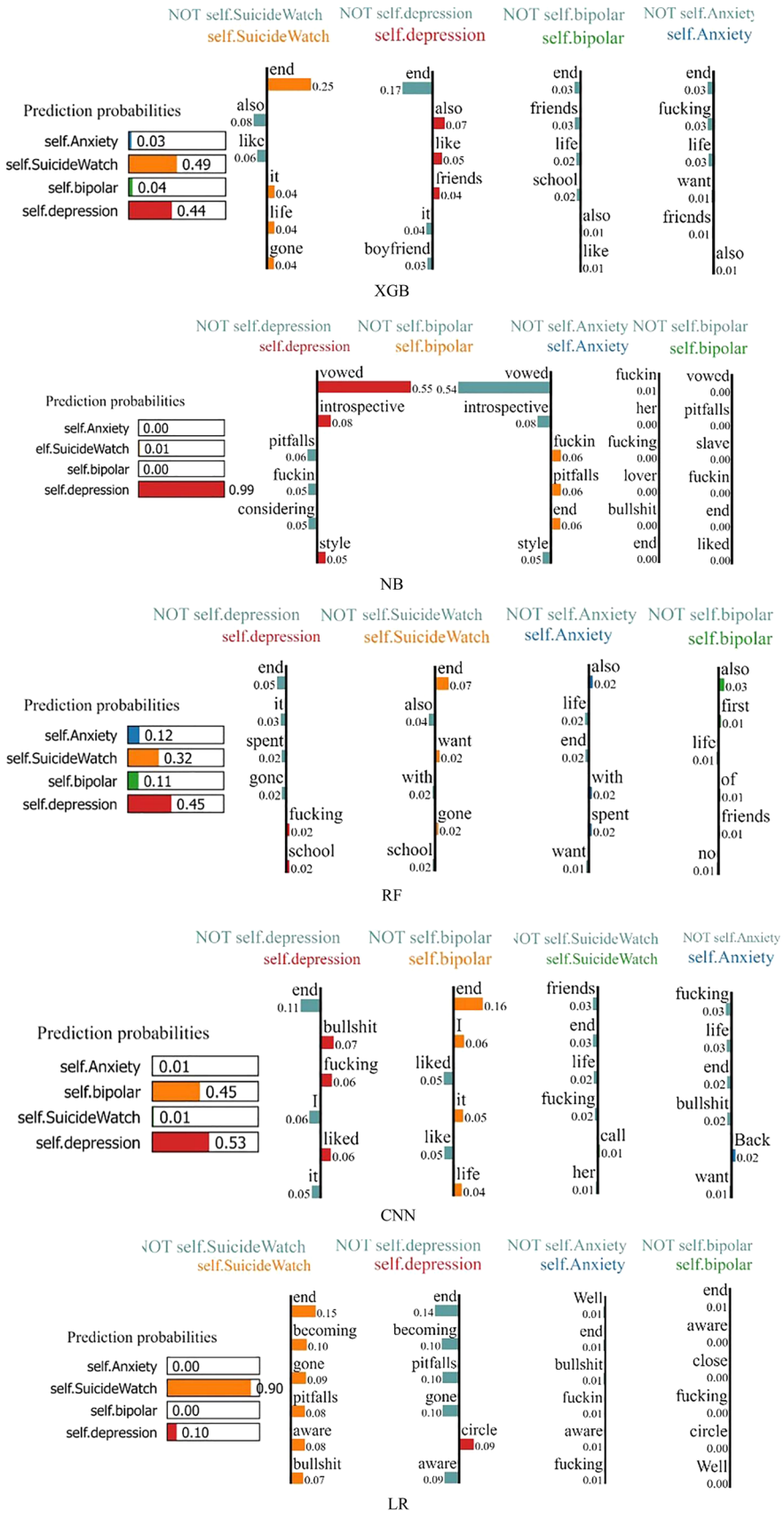


Table 4 (continued)

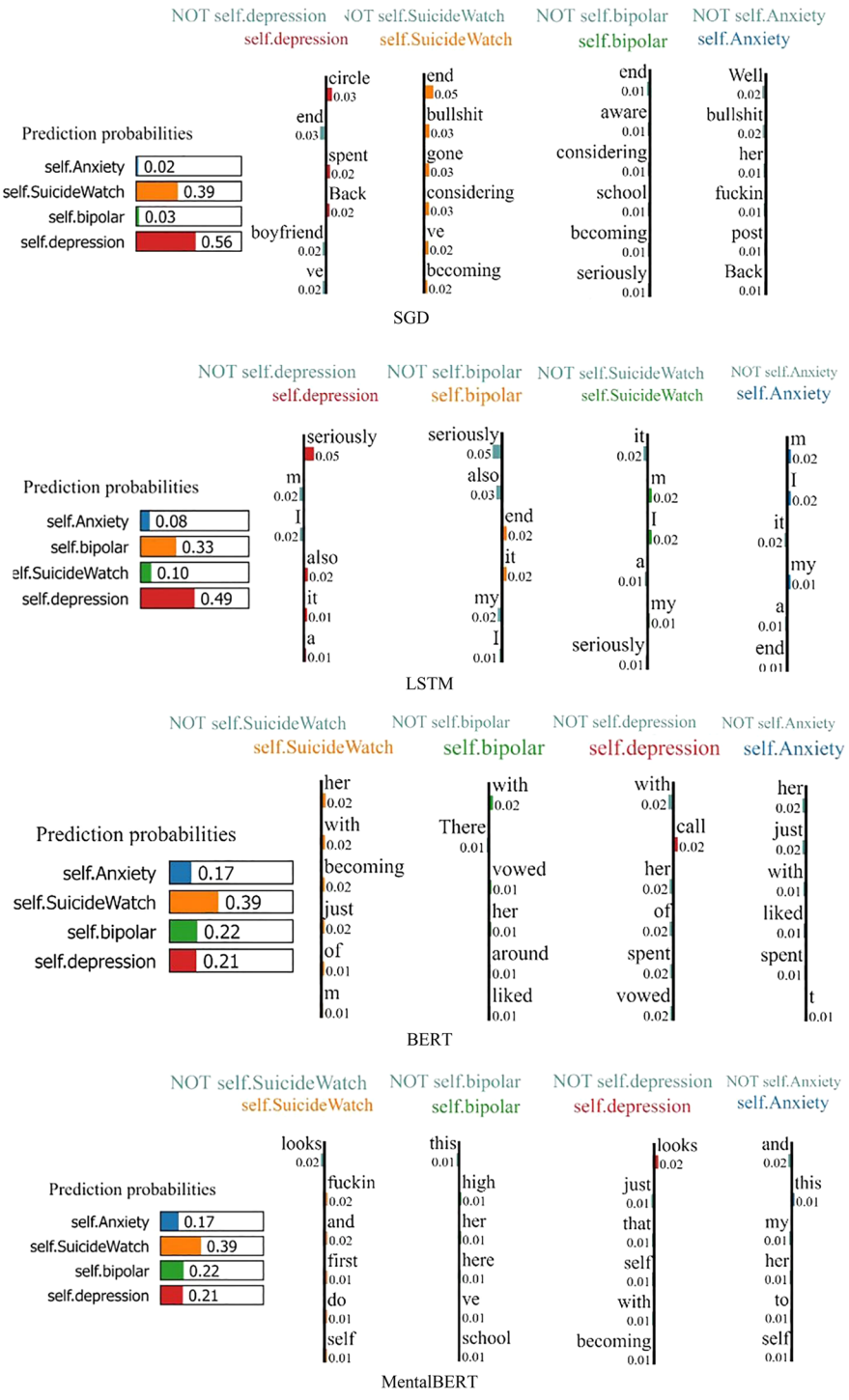


Table 5 Lime explanation for sample text C of each model

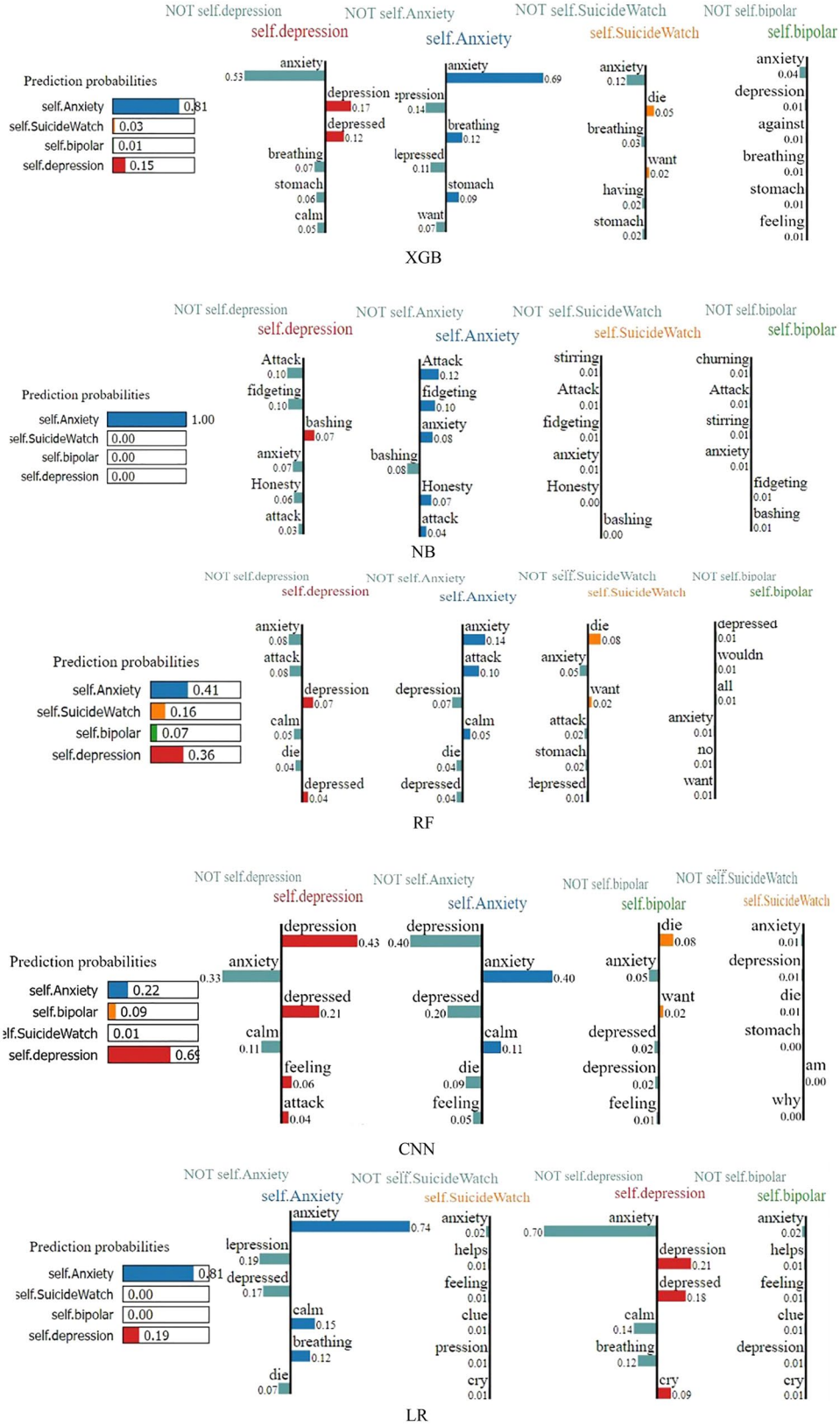
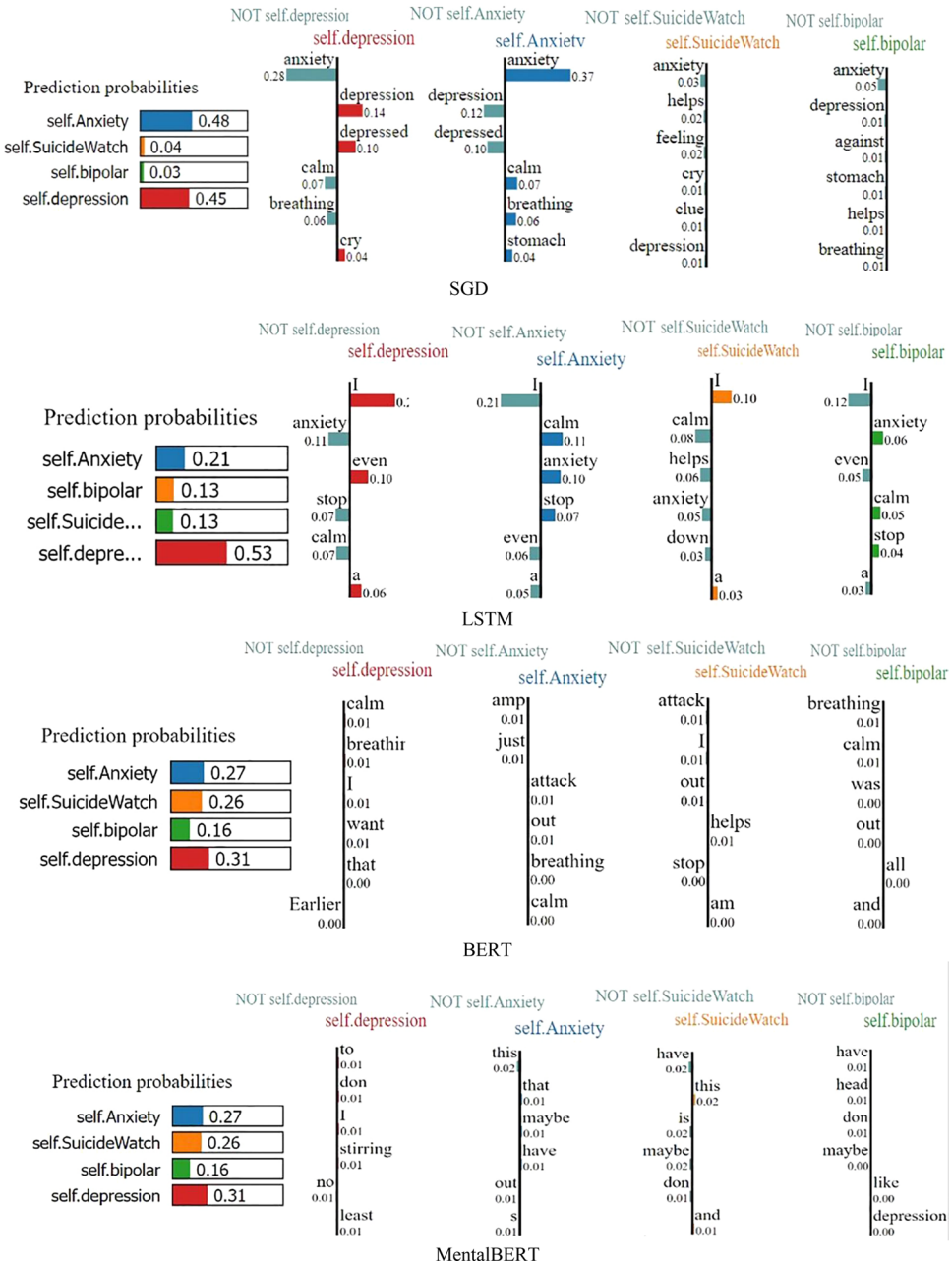


Table 5 (continued)



depression related words. Therefore, models XGB, NB, RF, SGD and LR predicted class self.Anxiety (Table 5). However, deep learning and transformer models CNN, LSTM, BERT AND MentalBERT classified actual class.

Acknowledgements The authors would like to thank Project SAMARTH, an initiative of the Ministry of Education (MoE), Government of India, at the University of Delhi South Campus (UDSC), for their support.

Authors contributions PDT was contributed to methodology and writing—original draft, NA was contributed to conceptualization, methodology, visualization, validation, and writing—original draft, VV was contributed to methodology and writing—original draft, GJS was contributed to conceptualization, visualization, validation, and writing—review and editing, **SS** was contributed to writing—review and editing, AP was contributed to validation and writing—review and editing.

Funding No funding was received for conducting this study.

Data availability The dataset analyzed in the present study can be available on request using the link: https://doi.org/10.5281/zenodo.6476179.

Code availability The two models of BERT and MentalBERT for this dataset have been released in Hugging Face hub and they can be downloaded for further studies.

MentalBERT: https://huggingface.co/tiya1012/swmh4_mtb
BERT: https://huggingface.co/tiya1012/swmh4_bert

Declarations

Conflict of interests The authors have declared that they have no conflict of interest.

Ethics approval and consent to participate Not applicable.

References

Abuzayed A, Al-Khalifa H (2021) BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. Procedia Computer Science 189:191–194. https://doi.org/10.1016/j.procs.2021.05.096

Alotaibi W, Alomary F, Mokni R (2023) COVID-19 vaccine rejection causes based on Twitter people's opinions analysis using deep learning. Soc Netw Anal Min 13:62. https://doi.org/10.1007/s13278-023-01059-y

Benrouba F, Boudour R (2023) Emotional sentiment analysis of social media content for mental health safety. Soc Netw Anal Min 13:17. https://doi.org/10.1007/s13278-022-01000-9

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research, 993–1022.

Boettcher N (2021) Studies of Depression and Anxiety Using Reddit as a Data Source: Scoping Review. JMIR Ment Health 8(11):e29487. https://doi.org/10.2196/29487

Breiman L (2001) Random Forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324

Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. ArXiv DOI 10(1145/2939672):2939785

Dao B, Nguyen T, Venkatesh S, Phung D (2015) Nonparametric discovery of online mental health-related communities. In: IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, France, pp 1–10. https://doi.org/10.1109/DSAA.2015.7344841

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://doi.org/10.48550/arXiv.1810.04805

"depression", Kaggle.com, 2021, [online] Available: https://www.kaggle.com/datasets/sahasourav17/students-anxiety-and-depression-dataset.

Garg M, Saxena C, Krishnan V, Joshi R, Saha S, Mago V, Dorr BJ (2022) CAMS: an annotated corpus for causal analysis of mental health issues in social media posts. arXiv preprint. https://doi.org/10.48550/arXiv.2207.04674

Gemmell J, Isenegger K, Dong Y, Glaser E, Morain A (2019) Comparing Automatically Extracted Topics from Online Mental Health Disorder Forums. In:International Conference on Computational Science and Computational Intelligence (CSCI), pp 1347–1352. https://doi.org/10.1109/CSCI49370.2019.00252

Gkotsis G, Oellrich A, Velupillai S, Liakata M, Hubbard TJ, Dobson RJ, Dutta R (2017) Characterisation of mental health conditions in social media using Informed Deep Learning. Sci Rep 7:45141. https://doi.org/10.1038/srep45141

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang GZ (2019) XAI—Explainable artificial intelligence. Sci Robot,4(37). https://doi.org/10.1126/scirobotics.aay7120

Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. Curr Opin Behav Sci 18:43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

Hanna F, Barbui C, Dua T, Lora A, van Regteren AM, Saxena S (2018) Global mental health: how are we doing? World Psychiatry 17(3):367–368. https://doi.org/10.1002/wps.20572

Hassan MM, Khan MAR, Islam KK, Hassan MM, Rabbi MMF (2021) Depression Detection system with Statistical Analysis and Data Mining Approaches. In: International Conference on Science & Contemporary Technologies (ICSCT), Dhaka, Bangladesh, pp 1–6, https://doi.org/10.1109/ICSCT53883.2021.9642550.

Hu Y, Sokolova M (2021) Explainable multi-class classification of the camh covid-19 mental health data. arXiv preprint. https://doi.org/10.48550/arXiv.2105.13430

Huang X, Wang S, Zhang M, Hu T, Hohl A, She B, Gong X, Li J, Liu X, Gruebner O, Liu R, Li X, Liu Z, Ye X, Li Z (2022) Social media mining under the COVID-19 context: Progress, challenges, and opportunities. International Journal of Applied Earth Observation and Geoinformation: ITC Journal 113:102967. https://doi.org/10.1016/j.jag.2022.102967

Islam MR, Kabir MA, Ahmed A, Kamal ARM, Wang H, Ulhaq A (2018) Depression detection from social network data using machine learning techniques. Health Inf Sci Syst 6(1):8. https://doi.org/10.1007/s13755-018-0046-0

Ji S, Li X, Huang Z, Cambria E (2022) Suicidal ideation and mental disorder detection with attentive relation networks. Neural Comput Appl 34(13):10309–10319. https://doi.org/10.1007/s00521-021-06208-y

Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E (2021) MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. arXiv https://doi.org/10.48550/arXiv.2110.15621

Kamarudin NS, Beigi G, Liu H (2021) A study on Mental Health Discussion through Reddit. In: International conference on software engineering and computer systems and 4th international conference on computational science and information management, ICSECS-ICOCSIM. https://doi.org/10.1109/ICSECS52883.2021.00122

Kathy L, Agrawal A, Choudhary A (2015) Mining Social Media Streams to Improve Public Health Allergy Surveillance. In:IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15), pp-815–822. https://doi.org/10.1145/2808797.2808896

Kilbourne AM, Beck K, Spaeth-Rublee B, Ramanuj P, O'Brien RW, Tomoyasu N, Pincus HA (2018) Measuring and improving the quality of mental health cara global perspective. World Psychiatry 17(1):30–38. https://doi.org/10.1002/wps.20482

Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content on social media. Sci Rep 10:11846. https://doi.org/10.1038/s41598-020-68764-y

Kotenko I, Sharma Y, Branitskiy A (2021) Predicting the Mental State of the Social Network Users based on the Latent Dirichlet Allocation and fastText. In:IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp 191–195. https://doi.org/10.1109/IDAACS53288.2021.9661061

Lin YS, Tai LK, Chen AL (2023) The detection of mental health conditions by incorporating external knowledge. J Intell Inf Syst. https://doi.org/10.1007/s10844-022-00774-w

Liu D, Feng XL, Ahmed F, Shahid M, Guo J (2022) Detecting and measuring depression on social media using a machine learning approach: systematic review. JMIR Ment Health 9(3):e27244. https://doi.org/10.2196/27244

Molnar, C (2022) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2. https://christophm.github.io/interpretable-ml-book

Pranckevi˘cius T, Marcinkevi˘cius V (2017) Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic J Modern Comput 5(2):221

Qi Y, Shabrina Z (2023) Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach. Soc Netw Anal Min 13:31. https://doi.org/10.1007/s13278-023-01030-x

Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S (2021) Depression detection on reddit with an emotion-based attention network: algorithm development and validation. JMIR Med Inf 9(7):e28754. https://doi.org/10.2196/28754

Ribeiro MT, Singh S, Guestrin C (2016) Why Should I Trust You?: Explaining the Predictions of Any Classifier. arXiv preprint. https://doi.org/10.48550/arXiv.1602.04938

Rizvi STR, Ahmed S, Dengel A (2023) ACE 2.0: A Comprehensive tool for automatic extraction, analysis, and digital profiling of the researchers in scientific communities. Soc Netw Anal Min 13:81. https://doi.org/10.1007/s13278-023-01085-w

Saha B, Nguyen T, Phung D, Venkatesh S (2016) A framework for classifying online mental health-related communities with an interest in depression. IEEE J Biomed Health Inf 20(4):1008–1015. https://doi.org/10.1109/JBHI.2016.2543741

Sangaraju VR, Bolla BK, Nayak DK, Kh J (2022) Topic modelling on consumer financial protection bureau data: an approach using BERT based embeddings. arXiv preprint. https://doi.org/10.48550/arXiv.2203.05794

Saxena C, Garg M, Ansari G (2022) Explainable causal analysis of mental health on social media data. Explainable causal analysis of mental health on social media data. arXiv preprint. https://doi.org/10.48550/arXiv.2210.08430

Stein DJ, Palk AC, Kendler KS (2021) What is a mental disorder? An exemplar focused approach. Psychol Med 51(6):894–901. https://doi.org/10.1017/S0033291721001185

Suicide data: Mental Health and Substance Use (2021). https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data Accessed 5 January 2023.

Verma R, Chhabra A, Gupta A (2023) A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective. Soc Netw Anal Min 13:12. https://doi.org/10.1007/s13278-022-01015-2

Wainberg ML, Scorza P, Shultz JM et al (2017) Challenges and opportunities in global mental health: a research-to-practice perspective. Curr Psychiatry Rep 19(5):28. https://doi.org/10.1007/s11920-017-0780-z

World mental health report: Transforming mental health for all - executive summary (2022). https://www.who.int/publications/i/item/9789240049338 Accessed 28 December2022.

Yazdavar AH, Mahdavinejad MS, Bajaj G, Thirunarayan K, Pathak J, Sheth A (2018) Mental health analysis via social media data. In: IEEE international conference on healthcare informatics (ICHI), NY, USA, pp 459–460. https://doi.org/10.1109/ICHI.2018.00102

Zanwar S, Wiechmann D, Qiao Y, Kerz E (2022) Exploring Hybrid and Ensemble Models for Multiclass Prediction of Mental Health Status on Social Media. arXiv preprint. https://doi.org/10.48550/arXiv.2212.09839

Zhou J, Zogan H, Yang S, Jameel S, Xu G, Chen F (2021) Detecting community depression dynamics due to covid-19 pandemic in Australia. IEEE Trans Comput Soc Syst 8(4):982–991. https://doi.org/10.1109/TCSS.2020.3047604