The Journal of Supercomputing
https://doi.org/10.1007/s11227-021-04040-8

Automatic detection of depression symptoms in twitter using multimodal analysis

Ramin Safa[1] · **Peyman Bayat**[1] · **Leila Moghtader**[2]

Accepted: 19 August 2021

Abstract

Depression is the most prevalent mental disorder that can lead to suicide. Due to the tendency of people to share their thoughts on social platforms, social data contain valuable information that can be used to identify user's psychological states. In this paper, we provide an automated approach to collect and evaluate tweets based on self-reported statements and present a novel multimodal framework to predict depression symptoms from user profiles. We used n-gram language models, LIWC dictionaries, automatic image tagging, and bag-of-visual-words. We consider the correlation-based feature selection and nine different classifiers with standard evaluation metrics to assess the effectiveness of the method. Based on the analysis, the tweets and bio-text alone showed 91% and 83% accuracy in predicting depressive symptoms, respectively, which seems to be an acceptable result. We also believe performance improvements can be achieved by limiting the user domain or presence of clinical information.

Keywords Mental health · Depression detection · Social media · Multimodal framework · Text mining

✉ Peyman Bayat
bayat@iaurasht.ac.ir

Ramin Safa
r.safa@outlook.com

Leila Moghtader
moghtaderleila@yahoo.com

[1] Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran

[2] Department of Psychology, Rasht Branch, Islamic Azad University, Rasht, Iran

Published online: 09 September 2021

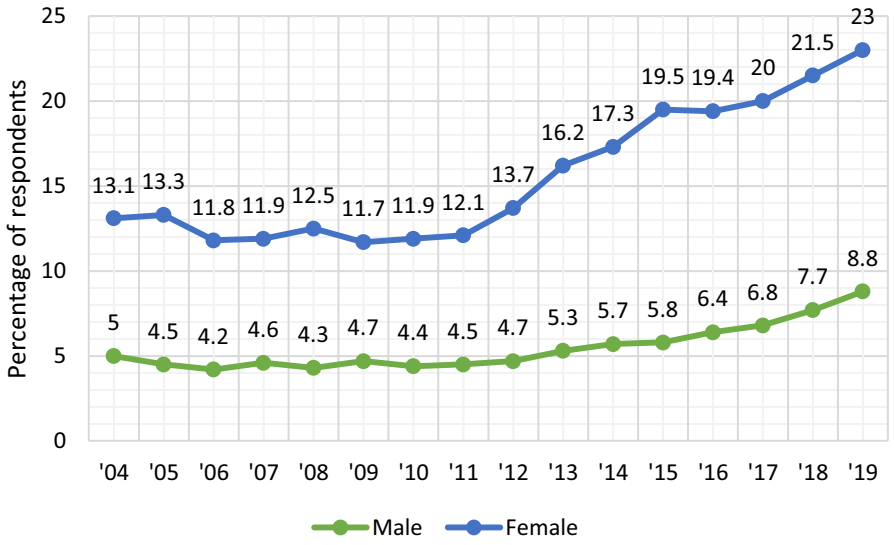


Fig. 1 The growing number of US youths with a major depressive episode from 2004 to 2019, by gender [6]

1 Introduction

In January 2020, the Mental Disorders Fact Sheet on World Health Organization (WHO) showed that, globally, more than 264 million people of all ages suffer from depression.[1] Recent findings also state that there is a high prevalence of mental health problems, during the COVID-19 outbreak [1]. While there are known, effective treatments for depression, only a few percentage people have received treatment for it [2]. The lack of appropriate treatment can lead to disability, psychotic episodes, thoughts of self-harm, and suicide, that is contributing to more than 800,000 deaths every year, and ranking as the second leading cause of deaths among 15 to 29 year olds [3, 4]. As an example of this trend, the percentage of US youths with a major depressive episode from 2004 to 2019, by gender is depicted in Fig. 1. Besides that, studies pointed out that the estimated economic value of mental illness is expected to reach 5 trillion dollars by 2030 [5]. It is clear that new prevention and intervention strategies are in high demand.

There is evidence that people increasingly turn to social media platforms such as Twitter and Facebook to represent their opinions, communicate with others, and share their feelings. This leads to big social data, containing traces of valuable information reflecting people's interests, moods, and behavior [2, 7, 8]. According to Hootsuite,[2] a well-known social media management platform, in July 2021, 4.48 billion people, or equal to almost 57% of the world's total population, are using

[1] https://www.who.int/news-room/fact-sheets/detail/depression
[2] https://blog.hootsuite.com/simon-kemp-social-media

social media. It is also stated that the global unique user total grew by 520 million over the past year, representing annual growth of more than 13%. These data provide a unique opportunity for researchers to understand users in detail. Conwey et al. highlighted that social media is established as a data source in various contexts, increasingly used in population health monitoring, and is beginning to be used for mental health applications [9]. While mental disorders are difficult to diagnose and monitor through traditional approaches, which heavily relying on surveys and interviews, online screening tools are valuable and might act in the future as more standard assessment strategies, like medical decision support systems [10] or health surveillance tools that can analyze signs of mental disorders. Predicting well-known symptoms might be done from user-generated content on social media, leading to new forms for the screening of the mental disorder. For example, many studies have highlighted that language patterns may serve as an indicator of the mental health state, also leading to the early detection of depression through machine learning techniques [2, 11–13].

The general process of using machine learning techniques in the research is as follows: 1) presenting a questionnaire for a predefined group of individuals, 2) request access to data and collection, 3) fitting the model based on the selected features and information extracted from the questionnaires, and 4) measuring the accuracy of estimation based on the test set. This is at a time when the lack of datasets is a major obstacle to the development of applied mental health research. Resolving this issue can be a key to crisis informatics [14], immediate diagnosis, intervention, and effective treatment. Questionnaires and surveys are the most widely used tools in the previous research, which are used to identify the psychological characteristics of the user. But the major problems with obtaining data through users' consent are few numbers of participants, significant cost, and a very time-consuming process [3, 15]. In addition, there are large temporal gaps in assessments of this approach; since identifying risk factors related to mental illness in many cases requires immediate intervention and it limits the development of effective intervention programs [16]. Moreover, people with a mental disorder may be less willing to cooperate with the research team. On the other hand, the reviewed works are mostly focused on textual features (specially tweets), while different types of social data are available that can be analyzed with the aim of investigating the psychological signals. For example, very few studies have examined visual features. This is despite the fact that the volume of images shared on social networks is very high; in some studies it has even been reported that images are the most important content shared on social platforms [17]. However, this is a fledgling research area and many details of the social data on mental disorders remain undiscovered.

In this paper, we first try to provide a road map for mental disorder prediction through related studies on social data mining and then present a new framework by automatically collecting the positive cases based on self-reported statements, examining the patterns of language use in tweets, bio-description, profile picture, and header image via lexicon analysis for detection of depression symptoms. To the best of our knowledge, this is one of the first efforts for automatically collecting large samples of depression symptoms from social media and the first study utilizing bio-text, mapping visual words to predefined lexicon, and utilizing profile header as a

feature on predicting mental disorders. Therefore, our main research questions are: 1) Is the presented framework a practical method (for field application) and confirms past findings? 2) Are there meaningful signals of depression in lexicon analysis of user-generated content? 3) Is it possible to achieve acceptable accuracy in predicting depression with the novel features? and 4) What features play a key role in the detection of depression? We try to test the performance of the proposed model for predicting potential depressed users. The major contributions of the study include:

- Providing an automated approach to collect and evaluate tweets based on self-reported statements and preparing a dataset to advance research in the field
- Presenting a multimodal framework to predict depression symptoms in Twitter users with in-depth analysis of the novel textual and visual features, from the lexicon perspective
- Developing several models for predicting depressive symptoms based on the selected features through correlation analysis and SVD

We believe that our multimodal framework can help identify potential sufferers with depression based on their profile information (not only the tweets). This study further argues that the proposed method can be used as a complementary tool to monitor the mental health status of individuals who use Twitter frequently.

The rest of the paper is structured as follows. Section 2 summarizes the related work. Section 3 presents our framework and details of its components followed to answer the research questions. Section 4 describes the configuration and experimental results; we explain how the system developed and present the corresponding results of the analysis. Finally, research findings, along with recommendations for future work, are presented in the discussion and conclusion sections.

2 Background and related work

Studies in the field of mental disorder prediction from social media platforms can be categorized into two groups according to data collection method: 1) collecting data directly from users with their agreement, using surveys and data collection tools, and 2) extracting data from public posts through APIs (Application Programming Interfaces). The first approach included posting research information on crowdsourcing platforms or data donation websites such as OurDataHelps,[3] and inviting users to participate in the application by filling questionnaires and consent allowing collection of their social data [12, 18, 19]. Center for Epidemiologic Studies Depression Scale (CES-D), and Beck Depression Inventory (BDI) are the most well-known questionnaires to measure participants' levels of depression. While Suicide Probability Scale (SPS) and Satisfaction with Life Scale (SWLS) are relevant instruments used to detect suicidal ideation, and a measure of satisfaction with life and well-being, respectively [20]. Since APIs in social media platforms allow developers

[3] https://ourdatahelps.org

to access public data, in the second approach, related posts are collected using associated keywords/phrases or regular expressions. For instance, some studies use "suicide," "self-harm," "kill myself," and "want to die" as main queries for post-retrieval, and some others use "I was diagnosed with [disorder]" for different kinds of mental health problems [15, 21, 22]; using this type of regular expressions is recognized as self-reported diagnosis. Extracting data through APIs lead to a collection of posts that should be evaluated before analysis. In case of relevant keywords/phrases, negation of suicide ideation, discussion of suicide of other people, or the news or reports consider as irrelevant and removed from the collection. Likewise, in the case of self-report diagnosis, only posts that did not contain hypotheticals statements, negations, or quotes are selected as positive samples by human assessment.

Given the problems we have outlined in the Introduction, many researchers recommend using regular expressions, usually with the help of human annotators for validation. There is also another way of obtaining data, which is using predefined datasets such as myPersonality[4] project, Computational Linguistics and Clinical Psychology (CLPsych)[5] [21], and eRisk[6] workshops [23], which provide both a variety of psychometric test scores and users social data for academic purposes. Since predefined datasets provide limited information (only the posts), in the following, we will mainly focus on studies that used the regular expressions approach. In most studies, in order to preprocess the collected data before the actual analysis, each post was preprocessed by removing stop-words, retweets, hashtags, URLs, and lowercasing characters [24–26]. Emojis were also converted to ASCII to facilitate future analysis. After feature extraction, a subset of relevant features is selected by feature selection approaches to reduce the training time, ease of interpretation, improving the chances of generalization, and avoid overfitting. The most widely used machine learning methods for mental disorder prediction are support vector machine (SVM) along with different kernels like linear, and radial basis function (RBF) [7, 27–32], different types of regression, such as linear, log-linear, and logistic [15, 33–36], naïve Bayes [27, 31, 37], decision tree [31, 37], and random forest [37–39]. Deep learning approaches have also been investigated for detecting individuals suffering from depression [40, 41], or recognizing suicide-related psychiatric stressors [24]. Though feature normalization process and parameter tuning were not well described in most previous studies, after prediction, the evaluation mechanism is employed to assess the reliability of the classification model. Classification accuracy, confusion matrix, precision, recall, F1-score, and receiver operating characteristics (ROC) curve are the most reported metrics and visualization tools in the literature, which helps to examine the performance of the proposed models.

A few research has investigated the use of different machine learning methods to find the relevant features [37, 40]. It should be noted that the development of learning methods is not the main focus of this research, but the study of new features that can be used in training a classifier and lead to the discovery of hidden patterns and

[4] https://sites.google.com/michalkosinski.com/mypersonality
[5] https://clpsych.org
[6] https://erisk.irlab.org

relationships in the data. Most of the studies use textual contents and linguistic patterns to understand what features have the biggest impact on mental disorder prediction. For instance, Coppersmith et al. [15, 42] showed that using natural language processing (NLP) methods on social data, disclose insights into particular mental health disorders, such as post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD). They also highlighted that related patterns of language, using first person pronouns, anger words, and various negative emotions have a strong relation with mental disorders. Part of these results was obtained by Linguistic Inquiry and Word Count (LIWC) [43] analysis. The LIWC is a well-known text analysis application that is often used to obtain linguistic patterns in related studies [11, 22, 26, 44]. It is manually constructed by psychologists and equipped with a set of dictionaries that covers various psychologically meaningful categories. It could be used to extract potential signals from the textual content, such as personal pronouns, and positive/negative emotion. The OpinionFinder [45] and SentiStrength [46] were also popular sentiment analysis tools, frequently used in selected research papers for quantifying the sentiment of textual expressions [47, 48]. Topic modeling techniques such as Latent Dirichlet Allocation (LDA) [49] have also been employed as a part of the content analysis in several efforts, to reveal latent topics from user posts [50, 51].

Nevertheless, these research efforts are heavily relying on textual features, and few studies have incorporated image analysis techniques on user-generated content [52, 53]. Kang et al. [7] used color compositions and SIFT descriptors as visual features to extract emotional meanings from the posted image on Twitter. Reece et al. [54] focus on using hue, saturation, and brightness of the image as features to predict signs of depression in Instagram users. In more recent work, Sharath et al. [55] have shown that image features such as color, facial, aesthetics, and content, besides utilizing VGG-Net [56] image classifier can be used to predict depression. To show the research gap and justify the significance of this study, in addition to the above background, we summarized the latest studies in Table 1.

As it turned out, only one work attempted to automatically collect positive samples from the social network. None of the previous studies investigated the lexicon analysis of the visual features of user-generated content. User bio-description and header image have not been considered and the relationship between these features remains undiscovered. Furthermore, to the best of our knowledge, this is the first time that this information has been analyzed for clues to mental disorders. Besides that, more ML models will be applied in our study that their ability to predict depression has not been previously investigated. To measure the performance of each modal in practice, all the evaluation metrics will be used to have a comprehensive assessment.

3 Methodology

The high-level architecture of the proposed framework is illustrated in Fig. 2; it is basically composed of three main modules. The first module is data collection and dataset building, which is a process of gathering depression diagnosed tweets and

Table 1 Summaries of some recent important studies

First author, date, reference	Aims	Collection method	Studied feature(s)	ML model(s) / Approach(es)	Metrics	SM platform
Our work	Presenting a multimodal automated framework for predicting potential depressed users from profile information	Self-report statements	Textual tweets, bio-text, profile picture, and banner image	SVM, LR, DT, Gradient Boosting, RF, Ridge-Classifier, AdaBoost, Catboost, and Multilayer Perceptron	Precision, Recall, F1-score, Accuracy, and AUC	Twitter
Zhou, 2021, [64]	Studying community depression dynamics due to COVID-19 pandemic in Australia	Self-report statements according to a specific location	Textual tweets	LR, Linear Discriminant Analysis, and NB	Precision, Recall, F1-score, Accuracy	Twitter
Ríssola, 2020, [3]	Presenting a textual dataset on automatic detection of depression	CLEF eRisk 2018 dataset	Textual posts	LR	Precision, Recall F1-score, and AUC	Reddit
Kim, 2020, [65]	Developing a deep learning model to identify user's mental state	Manual analysis and characterization of subreddits	Textual posts	XGBoost and Convolutional Neural Network	Precision, Recall, F1-score, Accuracy	Reddit
Guntuku, 2019, [66]	Finding which attributes of profile and posted images are associated with depression and anxiety	Survey-reported depression and anxiety (BDI and BAI)	Profile picture and posted images	Pearson Correlation	N/A	Twitter
Tadesse, 2019, [67]	Examining users' posts to detect factors that may reveal the depression	Manual analysis and characterization of subreddits	Textual posts	LR, SVM, AdaBoost, RF, Multilayer Perceptron	Precision, Recall, F1-score, Accuracy	Reddit
Islam, 2018, [68]	Aiming to perform depression analysis on Facebook data	Depression/non-depression indicative comments	Textual posts	DT, KNN, SVM, and Ensemble	Precision, Recall F1-score	Facebook

Table 1 (continued)

First author, date, reference	Aims	Collection method	Studied feature(s)	ML model(s) / Approach(es)	Metrics	SM platform
Ferwerda, 2018, [69]	Finding the relationship between the content of the uploaded Instagram pictures and the personality traits of users	Survey-reported personality traits (BFI)	Photos	K-means Clustering and Spearman's Correlation Analysis	N/A	Instagram
Chen, 2018, [70]	Employing fine-grained emotions as features in the task of identifying users with bipolar, depression, PTSD, and SAD disorder	Self-report statements	Textual tweets	RF, SVM, NB, LR, and DT	Precision, Recall, F1-score, and Accuracy	Twitter

RF = Random Forest, SVM = Support Vector Machine, NB = Naïve Bayes, LR = Logistic Regression, DT = Decision Tree, AUC = Area Under the ROC Curve

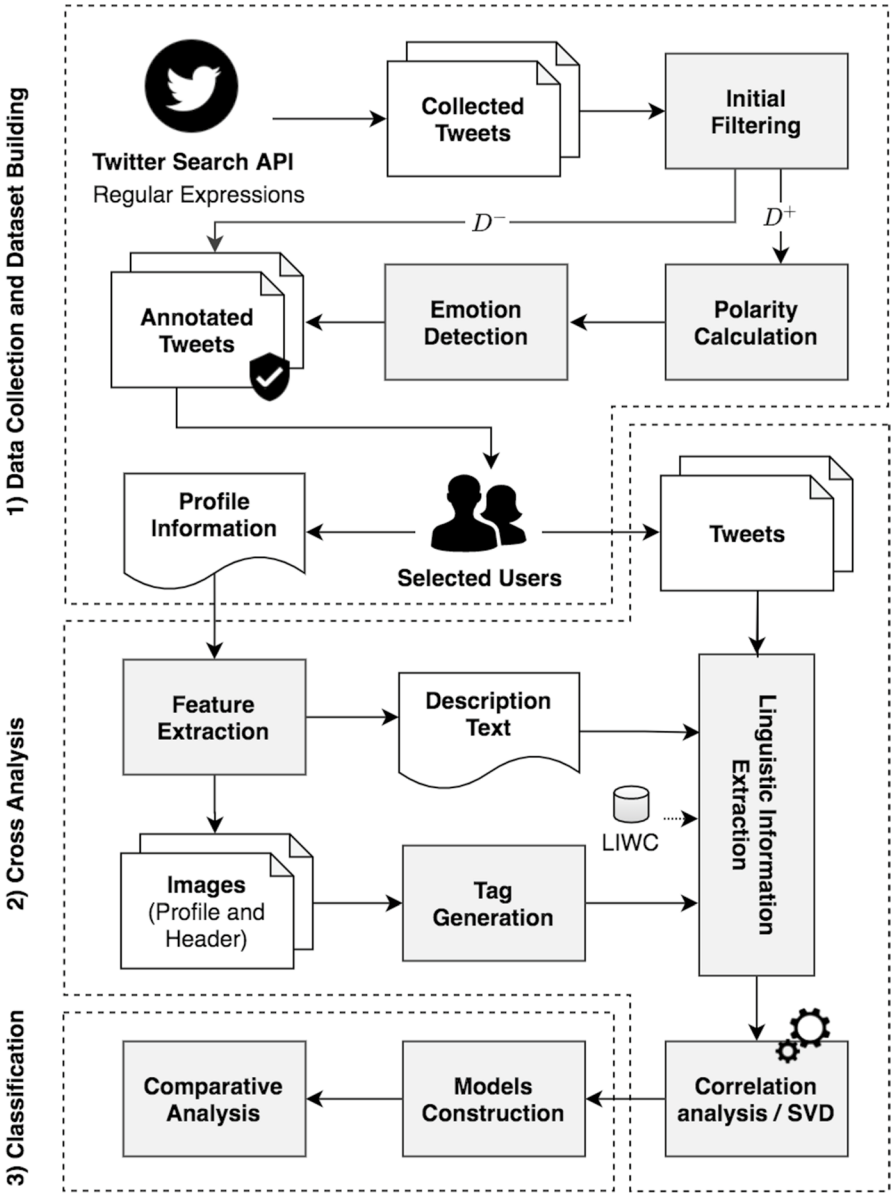


Fig. 2 The high-level architecture of the proposed framework consists of three main modules: 1) data collection and dataset building, 2) cross-analysis, and 3) classification

an automatic preprocessing pipeline to perform downstream analysis. The second module concerns the extraction of relevant features, and the cross-analysis of textual and visual features to find the effective ones. Finally, the third module presents the classification task to determine the user's psychological states, besides comparative

analysis. The theoretical details and implementation methods of each module are discussed in the sub-Sects. 3.1 to 3.3. The implementation of each module will also be presented in Sect. 4.

3.1 Data collection and dataset building

In order to identify the characteristics of online users with depression, we used Twitter as a data source. Since many users tend to publicly disclose information about their mental state, we first collected tweets with self-reported diagnosis using the regular expression "I was [just]/have been diagnosed with depression" with the Twitter API. As we explained in the previous section, it is required to ensure that the collected tweets have a genuine report about the diagnosis, and users were truly suffering from mental health conditions. Manual labeling by human annotators, including clinicians or reliable crowdsourcing workers, would be the ideal option, but it is very time-consuming. However, some preprocessing steps can significantly help to achieve the target data. Rissola et al. recently presented a method for automatically gathering post-samples of depression on Reddit[7] [3]; they proposed two heuristics to filter out less useful messages and characterize depression signs. The first heuristics method considered sentiment polarity score, and the second used topical similarity with a depression taxonomy. The obtained results showed that the first method has acceptable accuracy. Therefore, we follow their approach for building a dataset with some modifications as follows.

3.1.1 Initial filtering

At the first stage, several preprocessing steps need to be done to reach the target data. After collecting the tweets, all retweets were removed, because they are often an indication of someone else's post, that is not originally generated by the user. The duplicates and tweets that contained an URL were also eliminated, since we found most of them discussing news or advertisements instead of personal experience and ideation, and do not infer any useful information. Given that our method relies on the user-generated content and user profile data, we have to ensure that the necessary information is available, and the collected tweets were suitable to be analyzed. Hence, users who often post in non-English languages, or had less than 25 tweets, or don't provide the bio-information, profile picture, and header image, do not meet the requirements of this study and were excluded. The output of this initial filtering will be sets of tweets that have the potential to be analyzed but are still not sufficient for the application.

[7] https://reddit.com

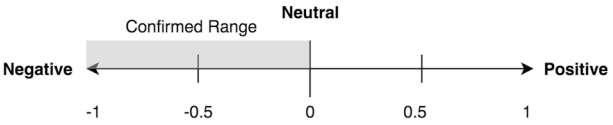


Fig. 3 The confirmed range of the polarity score, after initial filtering

3.1.2 Polarity calculation

The sentiment polarity score is calculated using a lexicon-based approach. The score ranges from -1 to 1, whereby less or equal to 0 values can indicate feelings of unhappiness or distress, especially when the tweets are written by users experiencing depression. As a result, it can be used as a measure for filtering candidate tweets. This idea is also used by Rissola et al. [3] with TextBlob[8] python library; but based on the reported findings [57] and our experimental analysis, Valence Aware Dictionary for Sentiment Reasoning (VADER)[9] can perform better than TextBlob in classifying tweets, since it is specifically attuned to sentiments expressed in social media. Hence, in this step, we use VADER to obtain the polarity score of each tweet and only keep those that have a value less or equal to 0 (Fig. 3).

3.1.3 Emotion detection

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [58], depressive moods are associated with the predominance of sadness and disgust. Consequently, if we can recognize these two emotions in the collected tweets, we will have a more reliable dataset to analyze. To this end, we utilize the NRC Emotion Lexicon (EmoLex) [59], which specifies the associations between a list of English words and eight basic emotions (fear, anticipation, trust, joy, anger, surprise, sadness, and disgust). So the tweets in the data set should have scored higher than a threshold value for sadness or disgust emotions. Rissola et al. [3] compute these scores by the average of the intensities of the words which evoke in emotion lexicon for each post, and used a threshold of 0.1 to filter out other posts; but as the tweets always have a word "depression" according to the input regular expression, this conditional phrase will always lead to a true conclusion. Thus, we modify the EmoLex lexicon by reducing the importance of the word "depression" on the "sadness" category, and based on empirical analysis, used a threshold of 0.5 for this emotion. The threshold of 0.1 was also used for the "disgust" emotion; however, fine-tuning the thresholds is left for future efforts. The summarized pseudocode of the first module is shown in Algorithm 1.

[8] https://textblob.readthedocs.io/en/dev
[9] https://github.com/cjhutto/vaderSentiment

Algorithm 1. Tweets validation and dataset building

Input

$Corpus$: Collected tweets using regex.

Output

$TList$: List of verified tweets (including user IDs).

$Corpus^{tmp} \leftarrow$ InitialFiltering($Corpus$)
// handling re-tweets, duplicates, URLs, and non-English tweets, besides filtering based on the number of tweets (at least 25), bio information, profile picture and, header
$TList \leftarrow \emptyset$
for each $Tweet \in Corpus^{tmp}$ **do**
  if PolarityScore($Tweet$) ≤ 0 **then**
    SScore, DScore $\leftarrow \emptyset$
    for each $Term \in$ Tokenization($Tweet$) **do**
      if SadnessScore(Term)$\neq 0$ **then**
        SScore $\longleftarrow SScore \cup \{SadnessScore(Term)\}$
      end if
      if DisgustScore(Term) $\neq 0$ **then**
        DScore $\longleftarrow DScore \cup \{DisgustScore(Term)\}$
      end if
    end for
    $\overline{SScore} \leftarrow \frac{1}{len(SScore)} \sum_{i=1}^{len(SScore)} SScore_i$
    $\overline{DScore} \leftarrow \frac{1}{len(DScore)} \sum_{j=1}^{len(DScore)} DScore_j$
    if $\overline{SScore} \geq 0.5$ or $\overline{DScore} \geq 0.1$ **then**
      $TList \longleftarrow TList \cup \{Tweet\}$
    end if
  end if
end for
return $TList$

After the above process, we will have a set of self-declared tweets from the candidate users. The users who posted these diagnosis statements were considered as potential candidates to form the diagnosed groups (D^+). In the next step, the required information of these users is collected by their ID. In a similar manner, to select a sample of users who do not suffer from depression and representing the general population, we collected the tweets, containing the keyword "the" using Twitter API (D^-) and considered the corresponding users as candidates of the control group (D_u^-). The difference between the processing of D^- and D^+ is that the overlapped users ($D_u^+ \cap D_u^-$) were removed from D^- to make sure it did not interfere with the training process, and the steps of measuring polarity and emotion detection were not considered for this set.

3.2 Cross-analysis

In this module, by extracting the desired features, a harmonic analysis technique is applied to measure the correlation, and thus determine the key features that best differentiate the diagnosed group from the control group.

3.2.1 Feature extraction

The user's mood state can be analyzed by both the textual and visual cues, from user-generated content. The information retrieved using the user ID contains specific number of recent tweets and comprehensive profile information from which the required features can be extracted, including bio-text, profile picture, and header image. The number of retrieved tweets and the preprocessing steps will be explained in the configuration section, but the remaining two types of features are analyzed as follows.

3.2.2 Tag generation

The images extracted from the user profile are the profile picture and the header, which is a large banner image placed at the top of a profile. By analyzing these images, we intend to discover latent patterns of depression. According to our studies, this is the first attempt to investigate the role of profile header in mental disorder prediction. To represent image content, we used Imagga,[10] the convolutional neural network-based automatic tagging system, which was effectively used in prior studies [17, 55]. Imagga Tagging API returns for each image a set of tags along with a confidence score. We label both profile and header image and generate a Bag-of-Visual-Words (BoVW) for each image; we only considered the top-10 predicted tags, which are the most important ones according to the developers' recommendations. Furthermore, we eliminate the tags that occurred less than 20 times in the whole data set.

Meanwhile, to reveal the predominant tags of each group, we extended both the diagnosed and control sets by considering tags from $D^{-}_{tags} \cup D^{+}_{tags}$ to allow comparisons. In this way, after normalizing the values of the two sets by MinMax normalization (Eq. 1), by subtracting the two groups from each other, we revealed the distinct differences (Eq. 2).

$$\overline{value} = \frac{value - min(value)}{max(value) - min(value)} \tag{1}$$

$$Diff_{tag} = \overline{D}^{-}_{tag} - \overline{D}^{+}_{tag} \tag{2}$$

3.2.3 Linguistic information extraction

Based on the idea that the words a person uses reflect his/her thoughts, emotions, and mental state, we examined two types of linguistic features: the LIWC features and language models. We use LIWC both as part of the analysis and as a source of features. As we explained before, LIWC is a text analysis tool to evaluate psychological, cognitive, and structural components of a given text, which uses a dictionary

[10] https://docs.imagga.com

composed of words and their classified categories or subdictionaries. So each entry can define one or more categories. After the necessary preprocessing, we generate Bag-of-Words (BoW) from the bio-text and collected tweets and consider along with the BoVW from the previous section, as three inputs of this phase. Consequently, we will have the scores of the categories corresponding to each entry. We propose that BoW and BoVW have complementary strengths with LIWC analysis and expect to find a significant correlation between them, in terms of psychological signals.

In addition, we employ n-gram language models to estimate the probability of certain character and word sequences. Due to shortenings and spelling errors in social media texts (especially Twitter [60]), it is not possible to use traditional word-based approaches ideally. Thus, again after the necessary preprocessing (such as stop-word removal), we employ two language models: the character n-gram, contain 2 to 4-g, and the word n-gram, include unigrams and bigrams. We also use the tf-idf technique for extracting features and further use them in the classification module. The term frequency (tf) of the n-gram t in document d is computed as follows.

$$tf(t,d) = \frac{c_{t,d}}{\sum_k c_{t,d}} \tag{3}$$

where $c_{t,d}$ denotes the number of times that t appears in d and $\sum_k c_{t,d}$ indicates the total number of terms in d. The inverse document frequency (idf) is also calculated as follows.

$$idf(t,D) = \log\frac{D}{d_t} + 1 \tag{4}$$

in which D is the total number of documents in the document set (corpus), and d_t is the number of documents in the corpus that contain t. Finally, the tf-idf is calculated by multiplying Eq. 3 by Eq. 4.

$$tf - idf(t,d,D) = tf(t,d) \times idf(t,D) \tag{5}$$

3.2.4 Correlation analysis

In order to assess what features are important for modeling, we perform a correlation analysis between the features. The Pearson's statistical correlation is applied for two purposes: 1) finding the inner correlation between features (BoW and BoVW, LIWC respective features, plus the n-grams), and 2) analyzing the relationship between different types of features, to reduce the feature set and selecting the effective ones for the classification. Equation 6 shows the Pearson correlation coefficient formula.

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{x}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}} \tag{6}$$

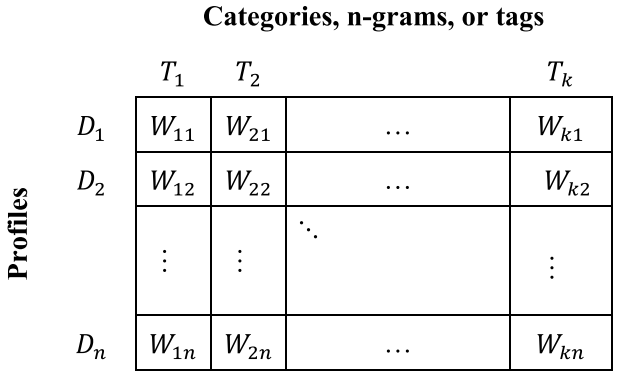


Fig. 4 The structure of the term-document in the case study (T = Term, D = Document, W = Weight)

In this formula, n indicates the sample size, x and y are the individual sample points indexed with i, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and analogously for $\bar{y}$. The larger the absolute value of r, the stronger the correlation between variables. We provide more details on this in the configuration section.

3.2.5 SVD approach

Singular value decomposition (SVD) is a dimensionality reduction technique that works by factorizing an *item*×*features* matrix A into three different matrices: an *item*×*concepts*, a *concept strength*, and a *concept*×*features* as represented by Eq. 7, where U and V are unitary orthogonal matrices, V^T is the transpose matrix of V, and λ is a nonnegative rectangular diagonal matrix.

$$A = U\lambda V^T \tag{7}$$

The most well-known application of SVD in NLP is latent semantic analysis, which is a theory and method for extracting and representing the meaning of words by statistical computations applied to a large corpus of text [61]. Latent semantic analysis can work with a term-document matrix that describes the occurrence of terms in documents. Since it can be used as a projection method where data with a large number of features are projected into a subspace with a smaller subset, while retaining the essence of the original data, we considered applying it to our problem as an alternative method. To reach this goal, we started with our special term-document matrix wherein each column represents the categories, n-grams, or tags which are the features extracted from the previous steps, and the rows represent one of the user profiles' main properties (ex. tweets). Each matrix entry indicates the normalized frequency of the corresponding term in the corresponding document (Fig. 4). We used truncated SVD to have a matrix with a lower rank and considered the outputs of this phase as inputs to train models.

3.3 Classification

We expect that the presented framework leverages the strengths of the selected features. In this module, using several supervised machine learning models, we examine different features in comprehensive analysis to predict the symptoms of depression disorder.

4 Configuration and Experimental Evaluation

In this section, first, we address data collection and the validation method for providing the dataset. Then, after in-depth feature analysis, nine classifiers were built to distinguish potential depress users from the control users; this will be done using a different combination of features including a novel feature set. Finally, the performance achieved by the classifiers was compared and evaluated.

4.1 The dataset

According to the methods that are explained in Sect. 3.1, we first collect the self-reported tweets using the regular expression "I was [just]/have been diagnosed with depression," by using the Twitter API. Based on this approach, we obtained diagnosis tweets, and consequently the primary diagnosed group. The duration of the collection process lasted about four months: from August 20, 2020, to December 9, 2020. In the next step, the initial filter was applied to prevent the analysis of disingenuous and misleading statements from these self-reported tweets. All retweets, duplicates, and tweets that contained an URL were removed from the collection. Also to ensure the availability of the required information, using the FastText[11] and the retrieved features, users who often post in non-English languages, or had less than 25 tweets, or doesn't provide the bio-description, profile picture and header were excluded from the analysis. Following Algorithm 1, and inspired by related studies, the top 553 tweets are selected to build the verified collection. Next, the corresponding user IDs are extracted from each tweet to form the candidate users (D_u^+). Each user ID refers to a specific user, and this will result in a diagnostic group of individuals. With a user ID, the crawler can access the user's public data. Thus, in the next step, user profile information including their bio-text, profile picture, and header image, in addition to 3,200 of their most recent tweets (according to the limitation of Twitter API) were downloaded; this results a total of 11,890,632 tweets, 1106 images (including profile picture and header image), and 553 bio-descriptions for analysis.

[11] https://github.com/facebookresearch/fastText

Table 2 Two samples of header images with their top-10 predicted tags

Labels	Header images	Top-10 Predicted tags
D^-		astronaut, aviator, man, person, male, people, helmet, professional, worker, happy
D^+		human, person, black, man, art, horror, male, cartoon, figure, skeleton

Table 3 Top-20 tags of profile picture and banner image from the diagnosed and control groups, sorted by difference absolute value

Feature Type	Distinct tags (sorted by difference absolute value)
Profile Picture	male, man, hair, pretty, attractive, symbol, sign, design, icon, face, portrait, people, graphic, cartoon, model, fashion, child, happy, eyes
Header Image	person, building, city, man, sea, water, architecture, symbol, texture, people, blank, sign, word, structure, ocean, border, sky, icon, happy

Similarly, we collected one day of tweets (December 16, 2020), containing the keyword "the," and considered the corresponding users as the candidate control group (D_u^-). We removed the overlapped users and skipped the steps of measuring polarity and emotion detection for this set. To create a balanced dataset, we chose a random sample of 570 users from the control group for the later classification experiments, and the result of this selection was 16,623,164 tweets, along with their 1140 images, and 580 bio-descriptions.

4.2 Feature analysis

All the extracted features need to be preprocessed before the analysis; this process includes the elimination of retweets, emoticons, URLs, various special characters, unicode characters, and mentioned users for textual data (tweets and bio). Also due to the fact that users sometimes post tweets in a non-English language, only English tweets were considered for analysis. In addition, for visual content, as Twitter allows the user to use a GIF format, GIF images were converted to JPG format for both profile and header images.

4.2.1 Image analysis

As it is explained in Sect. 3.2.2, we label both profile and header image with the Imagga API and considered the recommended top-10 predicted tags as a BoVW for each image. In Table 2, we visualize two sample header images from the diagnosed and control groups, along with their top-10 predicted tags.

The next step can be to find the co-occurred tags and consider only one of them to reduce the feature space; but we do not follow this path, as the remaining tags may not necessarily be semantically similar to each other and can help us to better interpret the upcoming results. Although to decrease sparsity, we remove the generated tags that occurred less or equal to 20 times in the whole data set, leaving us with 84 distinct tags for profile pictures and 111 for header images. The top-20 tags with the largest difference value between the two groups are shown in Table 3.

It is good to mention that regardless of the interpretation of the meaning of the tags in each group, the tags and their distribution can be an adequate measure for distinguishing these two groups, which we examine in the prediction section. But in the continuation of the proposed method, we will also analyze the above tags by mapping them to LIWC categories.

4.2.2 LIWC analysis

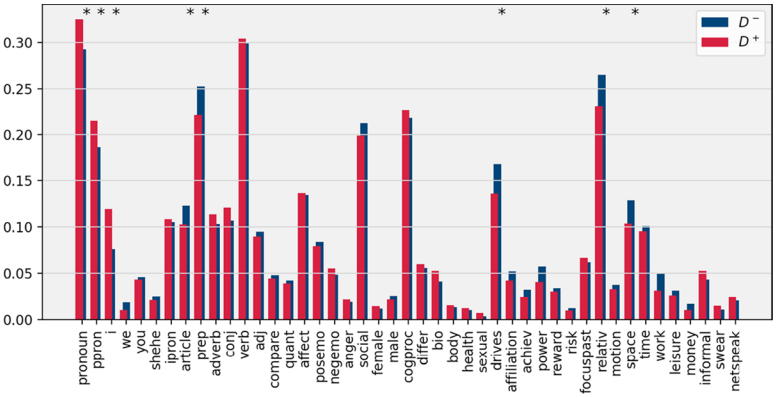
For the tweets and bio-texts, we use tokenization to extract tokens and generate the BoW. Then we used the generated BoWs and BoVWs to measure the proportion of the inputs that score positively on various LIWC categories. To perform our experiment, we extract determinant categories among 73 main LIWC2015 dictionaries from both the psycholinguistic point of view and the scores they achieved in each group, to convert the input features into numerical values.

Based on previous studies on depressed user tweets (most of them on LIWC2007) [15, 26, 27], the determinant LIWC categories include the first person pronoun, related patterns of language, varied negative emotions, as well as anger words compared to control users. In order to prove the proposed method and to ensure the correctness of the dataset, we replicate previous findings by revealing the differences in the distribution. For this purpose, after normalizing the scores of each category using MinMax normalization (according to Eq. 1), we identify and remove the categories in which the scores obtained show a slight difference between the two groups by Eq. 8; this will improve the interpretation of results. In Eq. 8, D^{-}_{tweets} and D^{+}_{tweets} represent control users' tweets and diagnosed users' tweets, respectively.

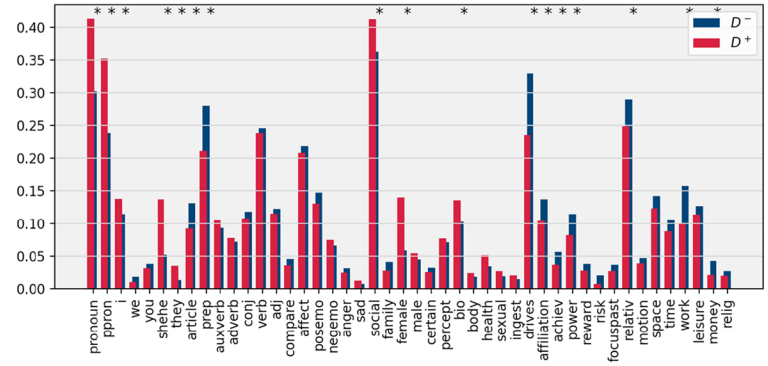
$$LIWC(Diff) = \overline{LIWC}\left(D^{-}_{tweets}\right) - \overline{LIWC}\left(D^{+}_{tweets}\right) \tag{8}$$

Finally, to better understand the differences in the distribution, we identified the distinct categories of the tweets by removing difference values smaller than the threshold of $\pm$ 0.002 and plotting the bar chart in Fig. 5a. The distinct categories (with an absolute difference higher than 0.02) are marked with an asterisk.

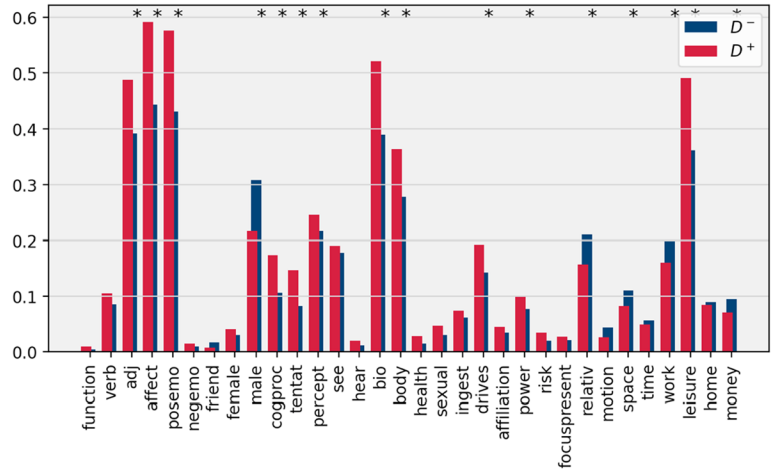
As can be seen in the chart, this distribution not only validates the collection method and the resulting dataset but also provides more details to expand discrimination. Moreover, for the first time, we examined the distribution of LIWC categories in the bio-text and the generated BoVWs from the profile picture and header image. The result of this experiment, after identifying the top differences (similar to Eq. 8, with a threshold of $\pm$ 0.005) is shown in Figs. 5b–d with normalized scores. A brief explanation of the definition of the selected categories is given in Table 4; but for further information, we refer the reader to the original source [43].



(a) Users' tweets

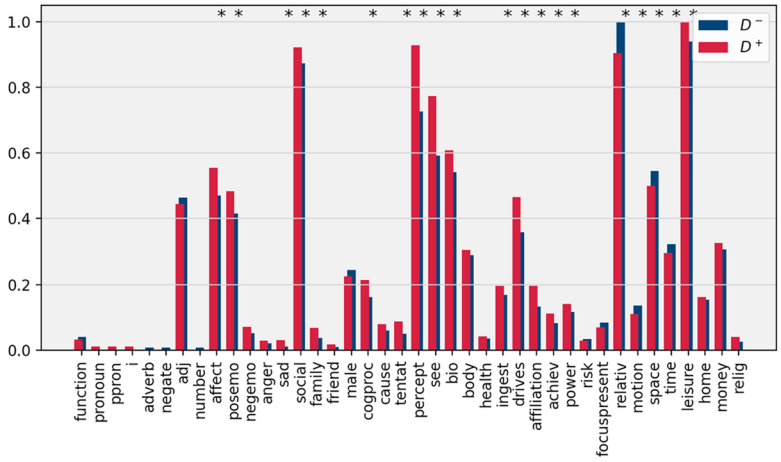


(b) Users' bi



(c) Users' profile pictures

Fig. 5 The proportion of distinct LIWC categories



(d) Users' profile banner

Fig. 5 (continued)

Table 4 Definition of some frequent LIWC features

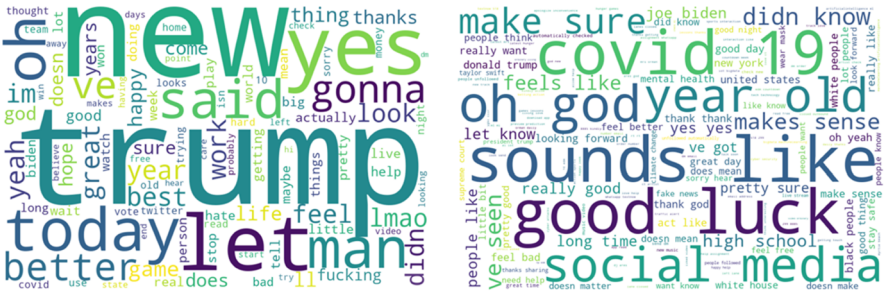
Feature	Definition	Examples
i	First person singular	I, me, mine
prep	Prepositions	to, with, above
drives	Drives and needs	Power, Risk focus
relativ	Relativity	area, bend, exit
social	Social processes	mate, talk, they
bio	Biological processes	eat, blood, pain
adj	Common adjectives	free, happy, long
affect	Affective processes	happy, cried
posemo	Positive emotion	love, nice, sweet
cogproc	Cognitive processes	cause, know, ought
sad	Sadness	crying, grief, sad
percept	Perceptual processes	look, heard, feeling
ingest	Ingestion	dish, eat, pizza
achiev	Achievement	win, success, better

4.2.3 N-gram Analysis

For n-gram modeling, first, we apply stop-word removal using the standard NLTK[12] English stop-words list. For better representation, we do not use lemmatization or stemming to group similar words together or convert words to their root form. We employ the character n-gram, contain 2 to 4-g, and the word n-gram, include

[12] https://nltk.org/nltk_data

(a) Diagnosed group



(b) Control group

Fig. 6 Unigrams and bigrams word cloud

unigrams and bigrams; then we use the tf-idf vectorizer from the scikit-learn python library[13] as a numeric statistic to highlight the importance of a term in each document of the corpus. Figure 6 presents the word cloud-based representation of the tweets' top-100 word unigrams and bigrams for both groups[14], in which the size of each word indicates its frequency in our experiment.

By examining the details, we observe evidence of hostility ("f**k," "sh*t," "hate"), sadness ("miss," "bad," "sorry"), sense of guilt ("feel bad," "im sorry"), self-oriented references and attention turned toward themselves ("im," "im just," "im gonna"), signs of help-seeking ("need help," "really need"), psychological statements ("mental health," "mental illness," "diagnosed depression"), and loneliness, with the tendency to their own feelings in the diagnosed group. In contrast to the diagnosed group, the n-grams examined in the control group contain mainly the words describing social and daily life, positive attitude, and feelings. The number of generated n-grams and the different types of methods for feature analysis are described in the third column of Table 5. We will explain the fourth and fifth

[13] https://scikit-learn.org

[14] In order to better illustrate the differences, top-25 frequent and similar n-grams have been removed from both groups.

Table 5 Different types of approaches to text encoding methods

Feature Type	Methods	# of Main Features	# of Selected Features via CA	# of Selected Features via SVD
LIWC	Tweet	73	22	15
	Bio-description	73	11	31
	Profile Picture	73	12	15
	Header Image	73	18	17
$\text{Word}_{1,2\,g}$	Tweet	3000, 3000	2942, 2610	590, 664
	Bio-description	3000, 3000	411, 1165	716, 735
$\text{Char}_{2,4\,g}$	Tweet	3000, 3000	2750, 2400	445, 497
	Bio-description	1398, 3000	875, 497	339, 618
Tagger	Profile Picture	84	23	44
	Header Image	111	30	62

Char = Character, CA. = Correlation analysis

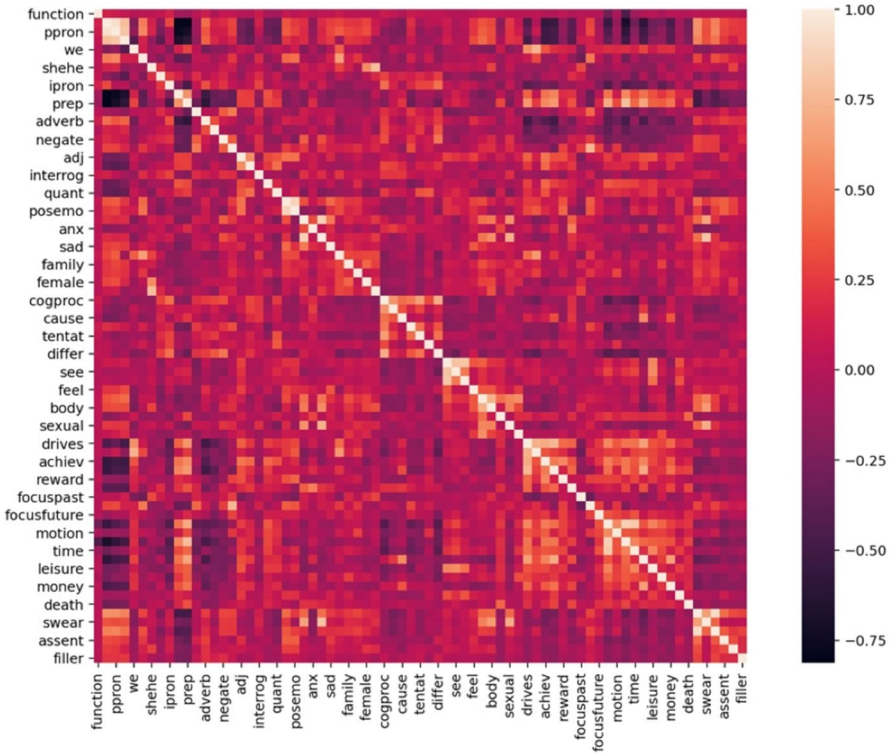


Fig. 7 The Pearson correlation heatmap among the LIWC dictionaries for both D^{+}_{tweets} and D^{-}_{tweets}

columns in the next section, where statistical correlations and SVD are applied for in-depth analysis of the features.

4.3 Correlation analysis

There are several types of correlation coefficients (e.g., Pearson, Spearman, and Kendall), and the most widely used is the Pearson's correlation coefficient. To analyze the relationship between the feature sets, Pearson's statistical correlations were extracted separately for each set, and an example of which is given in Fig. 7. This heatmap shows the correlation between LIWC dictionaries based on the score obtained by analyzing the tweets of the two groups.

As shown in the color bar, the bright tiles specify a positive correlation while the dark tiles indicate a negative correlation between two features, the shade of the colors indicates the strength from 100% positive to 100% negative correlations. Due to the high size of the features and the presence of annotation in each heatmap, we skip the visual representation of other correlations and explain how to use correlation analysis in selecting the effective features.

Features with high correlation are more linearly dependent, therefore have almost the same effect on the dependent variable. Hence, when two features have a significant correlation, we can drop one. Given the collected data and multimodal framework, at this stage, we do not provide a specific measure for selecting the target feature. But we point out that in the absence of a uniform distribution of feature values, frequency and availability are parameters that can influence the selection. So by empirical analysis, we consider the threshold to be 0.8 and we remove one of the two features that have a correlation higher than this value. As a result, each partial dataset will only have columns with a correlation of less than the threshold. Next, we use the p-value (probability value) to assess if the result of an experiment is statistically significant. Our null hypothesis is that the selected combination of independent variables does not have any effect on the dependent variable. So the p-value gives us the probability of finding an observation, under the assumption that the hypothesis is true. We build a small regression model to calculate the p values and use this probability (statistically significant) to accept or reject the hypothesis. In other words, as removing different features will have different effects on the p-value for the dataset, and a p-value less than or equal to 0.05 indicates strong evidence against the null hypothesis, we can measure the p-value in each scenario to decide whether to keep a feature or not. As the workflow is summarized in Algorithm 2, the result of each run will be a distinct number of feature sets for the modeling phase.

Algorithm 2. Feature engineering

Input

$Corpus_i$: The partial dataset (Different combinations).

Output

$\overline{Corpus_i}$: The trimmed dataset.

$Corpus^{tmp} \leftarrow Corpus$
for each $Feature_A, Feature_B \in Corpus_i$ **do**
  if $correlation(Feature_A, Feature_B) \geq 0.8$ **then**
    Drop $Feature_A$ from $Corpus_i^{tmp}$
  end if
end for
for $j = 0$ *to* $len(Features \in Corpus_i^{tmp})$ **do**
  $Regressor \leftarrow$ fit $regression(Value_{Label}, Value_{Features})$
  $Max \leftarrow \max(Regressor_{Features}^{PValue})$
  if $Max \geq 0.05$ **then**
    for $k = 0$ *to* $len\left(Features \in Corpus_i^{tmp}\right) - j$ **do**
      if $Regressor_{Feature_k}^{PValue} == Max$ **then**
        Drop $Feature_k$ from $Corpus_i^{tmp}$
        // consider new features set
      end if
    end for
  end if
end for
$\overline{Corpus_i} \leftarrow Corpus_i^{tmp}$
return $\overline{Corpus_i}$

Through the above process, we have identified the effective features, which can be expected to contain patterns and information of the user's depressive state. The number of selected features via correlation analysis are shown in Table 5. Also, to investigate the possible compounds for classification, different feature combinations were selected considering different accessibility scenarios (such as combining different textual/visual features together, and a combination of all type features), which we will explain in the Classification and Discussion section.

4.4 Classification

In order to evaluate the suggested method, in addition to a comparison of different features functionality, we created the benchmark classifier using a Logistic Regression and test a variety of different classifiers for the prediction tasks. We have applied nine classification methods, including the Decision Tree, Linear SVM, Gradient Boosting Classifier, Random Forest, RidgeClassifier, AdaBoost, Catboost, and Multilayer Perceptron (MLP). For MLP, we used the default settings in the scikit-learn package [62] and carefully tuned the key parameters to configure the number of hidden layers, the number of neurons, and the activation functions for each partial dataset to reach an acceptable accuracy. For the implementation of each classifier, we use tenfold cross-validation to verify the results. To evaluate the classification techniques, we apply the standard evaluation metrics, such as accuracy of estimations, and F1-score consisting of precision and recall, which relies on a confusion

Table 6 Performance metrics

	Relevant	Non-Relevant
Retrieved	TP (True Positive)	FN (False Negative)
Not Retrieved	FP (False Positive)	TN (True Negative)

matrix (Table 6) incorporating the information about each prediction outcome and is defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

$$Precision = \frac{TP}{TP + FP} \tag{10}$$

$$Recall = TPR = \frac{TP}{TP + FN} \tag{11}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

We have also used the AUC (area under the ROC curve) as another common measure of the predictive quality, which considers the probability of the predicted class. The ROC curve is plotted with TPR (true positive rate) against the FPR (false positive rate) that is calculated as follows.

$$Specificity = \frac{TN}{TN + FP} \tag{13}$$

$$FPR = 1 - Specificity \tag{14}$$

These evaluation metrics of the multimodal analysis are presented in Table 7 and the complementary Figs. 8, 9, 10, and 11, which allows us a deeper analysis of the outputs.

Given that accuracy is the most popular measure for evaluating classification and is used in most studies on depression detection, we interpret the results by plotting the maximum accuracy obtained in each feature during the experiment.

As can be seen in the comprehensive table and shown in Fig. 8, the best accuracy is achieved with the tweets, especially with the Catboost and Gradient Boosting algorithms resulting in 98% accuracy, followed by the bio-feature and MLP model (92%), profile picture with MLP text classifier (69%), and header image with Decision Tree (63%). Also, the result of the F1-score shows the same order. To take a closer look at the tweet features analysis, we consider Fig. 9, which plots the highest accuracy achieved in the cross-validation experiment.

As it is clear from the results, word bigram as a single feature outperforms other features and has 98% accuracy with Catboost and Gradient Boosting model. Word unigrams with 95% accuracy, LIWC and character four-grams with 81% accuracy, and character bigrams with 71% accuracy are in the next positions. So recent tweets retrieved from the users can strongly indicate their mood. But we found that in situations where tweets are not available for reasons such as limited access and few numbers, the user's public information can also play an effective role. It is very interesting that the bio-text can show signs of depression with 92% accuracy while being much shorter than the user's tweets and publicly available. To better understand the effect of this feature, we compared different analyzing methods in Fig. 10.

Accordingly, character four-gram with the highest accuracy of 92% using MLP can be an acceptable predictive feature. Word unigram with 82% accuracy using SVM, LIWC with 70% accuracy using Ridge classifier, char bigram with 68% accuracy using Catboost, and word bigram with 59% accuracy using AdaBoost and Ridge classifier are in the next positions. It was also observed that the profile picture and banner can play a complementary role in diagnosis. Eventually, considering that tweets word bigrams and bio-character four-grams are ones of the most powerful features for predicting depressive symptoms, their ROC curves of the target classifiers are also shown in Fig. 11a–b. ROC, as a probability curve, results in AUC representing the measure or degree of separability. Indeed, the higher the AUC, the better the model's classifying ability. This means the AUC near to 1 implies a supreme model with a great separability measure. The legends in each chart list the AUC values from different classifiers and indicates that our models perform reasonably well in separating the two target classes.

4.4.1 SVD Results

For a more comprehensive analysis, we repeated the above experiments using SVD (instead of correlation). To do this, we used the TruncatedSVD from the scikit-learn python library to fit the datasets. Due to the dimensions of the different features, we considered the percentage of variance explained by each of the selected components equal or greater than 0.9 to extract the output features. Thereafter, we added the corresponding class labels to each document and cross-validate the model with the aforementioned classifiers to explore the results.

Based on the experiments, latent semantic analysis with SVD could greatly reduce the feature space, as can be seen in Table 5, but the correlation analysis approach showed more accuracy and efficiency. For a more accurate comparison of this experiment and as an example of analysis output, the result of the highest accuracy achieved by tweets analyzing methods with tenfold cross-validation through SVD is shown in Fig. 12. In summary, SVM, RidgeClassifier, and LR performed better than other models. Also, the ROC diagrams obtained from the analysis of tweets word bigrams and bio-character four-grams based on the SVD

Table 7 Top-5 comparison of different methods using different features

Classifier	Feature	Methods	Prec	Recall	Acc	F1	AUC
Logistic Regression	T	LIWC	0.65	0.72	0.67	0.68	0.67
		Char2	0.66	0.62	0.65	0.64	0.65
		Char4	0.71	0.70	0.71	0.70	0.71
		Word1	0.74	0.76	0.75	0.75	0.75
		Word2	0.74	0.75	0.75	0.75	0.75
	B	LIWC	0.62	0.55	0.61	0.58	0.61
		Char2	0.58	0.58	0.58	0.58	0.58
		Char4	0.72	0.55	0.67	0.62	0.67
		Word1	0.70	0.44	0.63	0.54	0.63
		Word2	0.57	0.28	0.54	0.38	0.54
	P	LIWC	0.60	0.59	0.61	0.59	0.61
		Tags	0.56	0.55	0.57	0.56	0.57
	H	LIWC	0.51	0.46	0.52	0.48	0.51
		Tags	0.50	0.52	0.51	0.51	0.51
Gradient Boosting Classifier	T	LIWC	0.72	0.73	0.73	0.72	0.73
		Char2	0.65	0.62	0.65	0.63	0.65
		Char4	0.73	0.69	0.72	0.71	0.72
		Word1	0.89	0.86	0.88	0.87	0.88
		Word2	0.97	0.84	0.91	0.89	0.91
	B	LIWC	0.58	0.58	0.59	0.58	0.59
		Char2	0.59	0.58	0.60	0.59	0.60
		Char4	0.68	0.52	0.64	0.59	0.64
		Word1	0.67	0.36	0.60	0.47	0.59
		Word2	0.65	0.17	0.55	0.27	0.54
	P	LIWC	0.59	0.55	0.59	0.57	0.59
		Tags	0.57	0.44	0.56	0.50	0.56
	H	LIWC	0.50	0.41	0.50	0.45	0.50
		Tags	0.50	0.43	0.51	0.46	0.51
RidgeClassifier	T	LIWC	0.68	0.77	0.70	0.72	0.70
		Char2	0.64	0.63	0.64	0.63	0.64
		Char4	0.70	0.71	0.71	0.70	0.71
		Word1	0.74	0.73	0.74	0.73	0.74
		Word2	0.77	0.75	0.77	0.76	0.77
	B	LIWC	0.62	0.55	0.61	0.58	0.61
		Char2	0.56	0.58	0.57	0.57	0.57
		Char4	0.74	0.63	0.71	0.68	0.71
		Word1	0.70	0.50	0.65	0.58	0.64
		Word2	0.57	0.28	0.54	0.37	0.54
	P	LIWC	0.61	0.59	0.61	0.60	0.61
		Tags	0.58	0.56	0.58	0.57	0.58
	H	LIWC	0.52	0.46	0.52	0.49	0.52
		Tags	0.50	0.51	0.50	0.50	0.50

Table 7 (continued)

Classifier	Feature	Methods	Prec	Recall	Acc	F1	AUC
Catboost	T	LIWC	0.71	0.75	0.73	0.73	0.73
		Char2	0.65	0.63	0.65	0.64	0.65
		Char4	0.75	0.72	0.74	0.73	0.74
		Word1	0.89	0.86	0.88	0.87	0.88
		Word2	0.99	0.82	0.91	0.89	0.91
	B	LIWC	0.58	0.58	0.58	0.57	0.58
		Char2	0.59	0.58	0.59	0.58	0.59
		Char4	0.69	0.58	0.67	0.63	0.66
		Word1	0.65	0.40	0.60	0.50	0.60
		Word2	0.58	0.18	0.53	0.27	0.53
	P	LIWC	0.58	0.54	0.58	0.56	0.58
		Tags	0.57	0.48	0.57	0.52	0.57
	H	LIWC	0.51	0.43	0.51	0.47	0.51
		Tags	0.52	0.48	0.52	0.49	0.52
MLP	T	LIWC	0.71	0.77	0.73	0.74	0.73
		Char2	0.60	0.62	0.61	0.61	0.61
		Char4	0.68	0.69	0.68	0.68	0.68
		Word1	0.69	0.72	0.70	0.70	0.70
		Word2	0.69	0.72	0.70	0.70	0.70
	B	LIWC	0.61	0.54	0.60	0.57	0.60
		Char2	0.52	0.52	0.53	0.52	0.53
		Char4	0.86	0.79	0.83	0.82	0.83
		Word1	0.76	0.65	0.72	0.69	0.72
		Word2	0.51	0.79	0.52	0.62	0.53
	P	LIWC	0.60	0.59	0.61	0.60	0.61
		Tags	0.58	0.52	0.58	0.55	0.58
	H	LIWC	0.53	0.42	0.53	0.47	0.53
		Tags	0.51	0.51	0.52	0.51	0.52

Acc. = Accuracy, Prec. = Precision, T = Tweets, B = Bio-description, P = Profile picture, H = Header image

approach are presented in Fig. 13. Although the results obtained from the tweets word bigrams are debatable, still the correlation-based analysis yields more promising results.

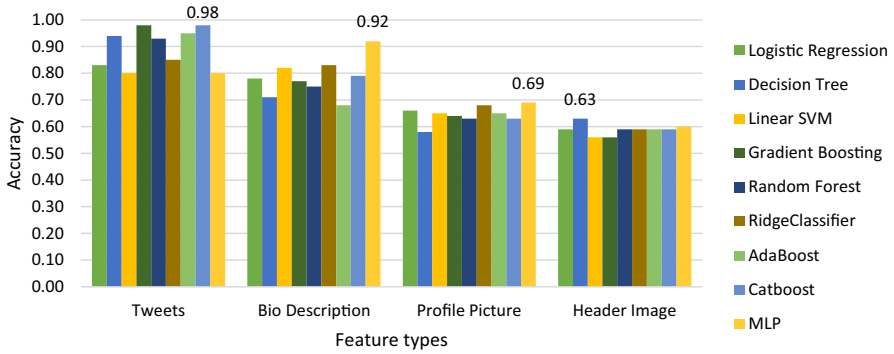


Fig. 8 Comparison of the highest accuracy achieved on feature types by nine different classifiers

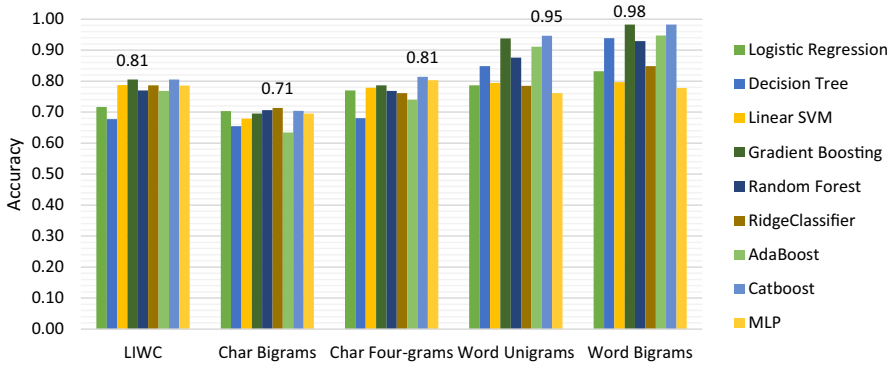


Fig. 9 Comparison of the highest accuracy achieved by tweets analyzing methods

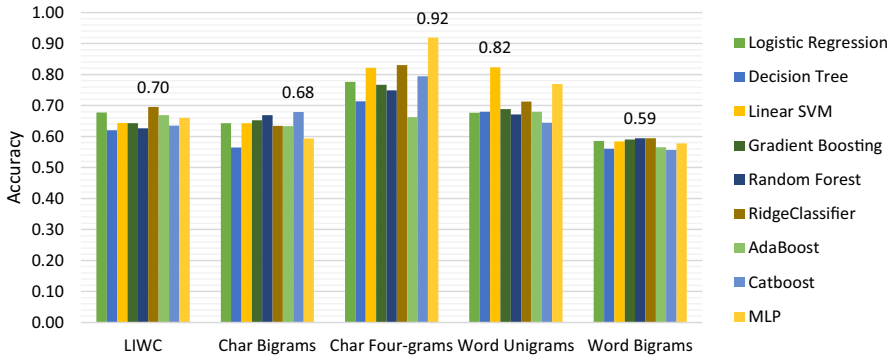
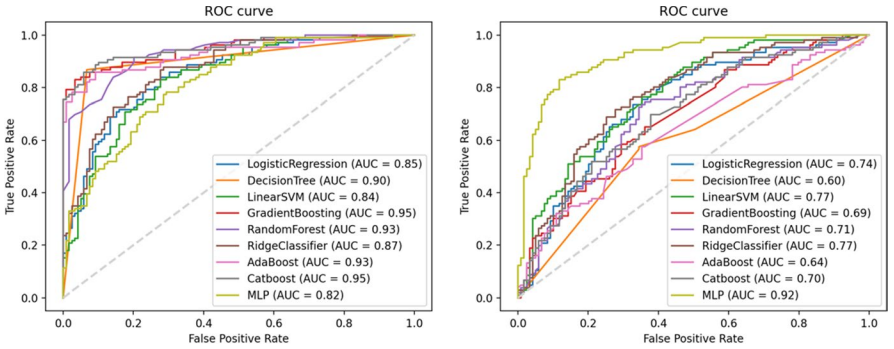


Fig. 10 Comparison of the highest accuracy achieved by bios analyzing methods



(a) Tweets.Word2 **(b)** Bio.Char4

Fig. 11 ROC curves of the target classifiers

5 Discussion

Although the proposed methods can be a proof-of-concept for similar implementations and have shown promising results, we believe that the results' accuracy could be even greater. In evaluation using real-world data, when the proposed method shows 91% accuracy, it means that based on the matching process, on average, in 91% of the cases the model has achieved 100% relevant results, which are the users who have self-report statements. This is in a situation where we did not limit the user domain to a small community (as was done in many studies on Reddit and Facebook). In addition, there were no other user's health records (e.g., personality traits) available for validation except user statements.

Also, we randomly sampled a total of 200 profiles (100 from each class) and asked three experts[15] to determine which profile could be considered as a reference to depression symptoms following the DSM-5 criteria. Each profile was assigned with one of the following labels: 1) no depression reference is expressed, 2) one or more depression references are expressed, and 3) unable to make a judgment. Then we chose the most agreed one. Accordingly, we found that in many retrieved cases the system retrieved profiles that might have been prone to depressive patterns of thinking but had not (yet) reported the self-statements of depression. This indicated that if we were less strict in the evaluation phase, the overall accuracy would be significantly improved.

On the other hand, the results of the analysis confirm that the automatic data collection can be a practical, cost-effective, and less time-consuming approach than the traditional ways. This method can also be used for similar issues (such as suicide prevention frameworks or investigation of other mental disorders) with a slight modification. We also point out that in some cases the combination of features will improve the evaluation results, but there has to be a trade-off between increasing metrics and the features

[15] One PhD student in computer engineering (who is working on computational linguistic) and two PhD students in psychology (who have experience in social media mining)—none co-authoring this paper.

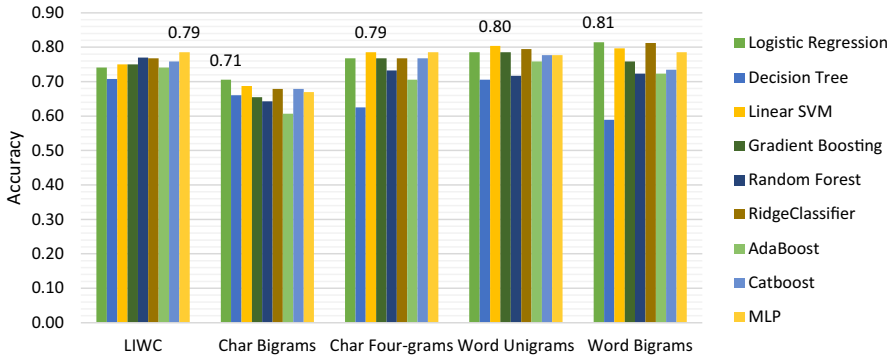


Fig. 12 Comparison of the highest accuracy achieved by tweets analyzing methods via SVD approach

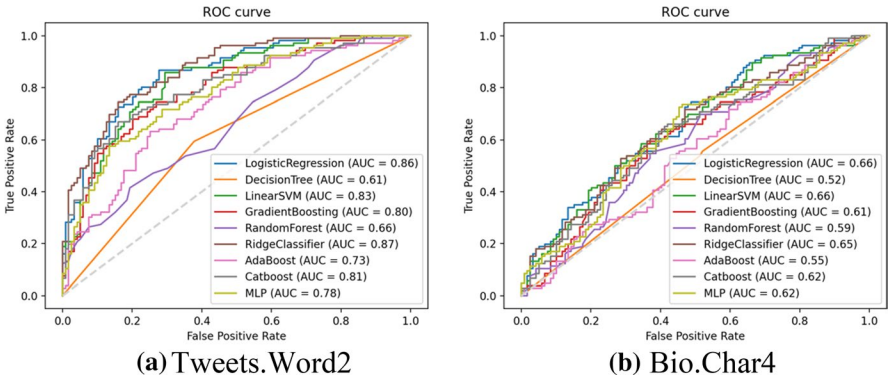


Fig. 13 ROC curves of the target classifiers for SVD analysis

involved. For example, after examining the cross-combinations, in some cases, combining three features leads to about a 1% increase in accuracy, which is not an optimal choice due to the feature dimensions, and using lightweight approaches (like what was described in the previous section) is a more appropriate choice.

6 Conclusions and future work

This study aims to detect depressed users on Twitter using self-report diagnosis. We first tried to provide a road map for mental disorder prediction via social data mining and then present a new framework by automatically collecting the data from the diagnosed and control user's profile. We extracted the desired features from the dataset and applied the preprocessing steps. Then we examined the distribution of data in defined categories to identify the differences between two groups and replicate previous findings. We tried to characterize the connection between depression and language use through lexicon analysis and NLP techniques, and we have

also introduced a set of new features that have not been further explored in previous research, including bio-text and features resulting from the analysis of user profile picture and banner image. We reduced the feature size to identify the effective ones by correlation analysis. Afterward, we applied nine classification models on the features and compared the results using three evaluation metrics: accuracy, F1-score, and area under the ROC curve. Thus, our findings suggest a relationship between depression symptoms and almost all the studied features, but tweets word bigrams and bio-character four-grams were identified as two important ones that in cross-validation using Catboost/GB model and MLP achieved the accuracy of 91% and 83%, respectively. Also, the F1-score of these two features was 0.89 and 0.82, which outperform the reported results with LR and SVM (common classifiers in the literature). Also in the case of dimensionality reduction techniques, in an alternative approach, we used SVD for feature selection, which led to the smaller set of features, and by comparing the implementation results, we showed that the correlation-based method leads to better outcomes. Furthermore, due to the importance of the interpretation of features in this study, the correlation-based method was a more proper choice.

We believe that the proposed mechanism can be implemented for other mental disorders with a slight change in the initial filtering phase. This framework can also be used as a lightweight method in the form of clinical decision support systems to facilitate diagnosis decisions or as suicide and self-harm prevention tools in social network platforms. We mentioned in the discussion section that if we were less strict in the evaluation, the results could be improved. In addition, although we tried to use all the important features in the user profile, there is other information that we believe can increase the accuracy of the diagnosis. This information includes hashtags and context information (e.g., tweets time, which showed a meaningful difference in a study on Sina Weibo [63]). For this reason, we also stored this information in the generated dataset to be used for future studies.

Finally, it is good to point out that behavioral/derived data are not limited to reported statements and profile information but includes all the activities and actions performed by the user. Such as the content that the user normally follows the games that he plays on the social network, or even the time he spends on the platform. In the presence of clinical information and patient approval, some of this information can be explicitly recorded and collected by the API or the crawler, and others implicitly by developing the application on the platform or installing add-ons in the user's browser. The development of such tools and the study of the collected data can create a new generation of analyses and open new gates to social network analysis in the field of diagnosing the user's mental state. Accordingly, the use of clinical information, semantic similarity techniques, transformer-based machine learning models, and more image analysis approaches are among the issues that will be addressed in our future work.

References

1. Gao J et al (2020) Mental health problems and social media exposure during COVID-19 outbreak. PLoS ONE 15(4):e0231924
2. Martínez-Castaño R, Pichel JC, Losada DE (2020) A big data platform for real time analysis of signs of depression in social media. Int J Environ Res Public Health 17(13):4752
3. Ríssola EA, Bahrainian SA, Crestani F. (2020) A Dataset for Research on Depression in Social Media. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 338–342.
4. James SL et al (2018) Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. The Lancet 392(10159):1789–1858
5. Bloom DE *et al.*, (2012) The global economic burden of noncommunicable diseases. Program on the Global Demography of Aging
6. SAMHSA (2021) Major depressive episode in the past year among U.S. youths by gender 2004–2019. Statista - The Statistics Portal. https://www.statista.com/statistics/252323/major-depressive-episode-among-us-youths-by-gender-since-2004/ Accessed 12 August 2021
7. Kang K, Yoon C, Kim EY (2016) Identifying depressive users in Twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, 2016: IEEE, pp. 231–238.
8. Javadi S, Safa R, Azizi M, Mirroshandel SA (2020) A Recommendation System for Finding Experts in Online Scientific Communities. J AI Data Min 8(4):573–584
9. Conway M, O'Connor D (2016) Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 9:77–82
10. Ebert DD, Harrer M, Apolinário-Hagen J, Baumeister H, (2019) Digital interventions for mental disorders: key features, efficacy, and potential for artificial intelligence applications. In *Frontiers in Psychiatry*: Springer. 583–627.
11. Loveys K, Crutchley P, Wyatt E, Coppersmith G, (2017) Small but mighty: affective micropatterns for quantifying mental health from social media language. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017, pp. 85–95.
12. Coppersmith G, Leary R, Crutchley P, Fine A (2018) Natural language processing of social media as screening for suicide risk. Biomed Inform Insights 10:1178222618792860
13. Plaza-del-Arco FM, Martín-Valdivia MT, Ureña-López LA, Mitkov R (2020) Improved emotion recognition in Spanish social media through incorporation of lexical knowledge. Futur Gener Comput Syst 110:1000–1008
14. Park A, Bowling J, Shaw G, Li C, Chen S (2019) Adopting social media for improving health: opportunities and challenges. N C Med J 80(4):240–243
15. Coppersmith G, Dredze M, Harman C (2014) Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2014, pp. 51–60.
16. De Choudhury M (2013) Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia.* pp. 49–52.
17. Samani ZR, Guntuku SC, Moghaddam ME, Preoţiuc-Pietro D, Ungar LH (2018) Cross-platform and cross-interaction study of user personality based on images on Twitter and Flickr. PLoS ONE 13(7):e0198660
18. Schwartz HA *et al.* (2016) Predicting individual well-being through the language of social media. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, 2016: World Scientific, pp. 516–527.
19. Braithwaite SR, Giraud-Carrier C, West J, Barnes MD, Hanson CL (2016) Validating machine learning algorithms for Twitter data against established measures of suicidality". JMIR Mental Health 3(2):e21
20. Wongkoblap A, Vadillo MA, Curcin V (2017) Researching mental health disorders in the era of social media: systematic review. J Med Internet Res 19(6):e228. https://doi.org/10.2196/jmir.7215

21. Losada DE, Crestani F (2016) A test collection for research on depression and language use. International Conference of the Cross-Language Evaluation Forum for European Languages. Springer, pp 28–39
22. Ríssola EA, Aliannejadi M, Crestani F (2020) Beyond Modelling: Understanding Mental Disorders in Online Social Media. European Conference on Information Retrieval. Springer, pp 296–310
23. Losada DE, Crestani F, Parapar J (2020) eRisk 2020: Self-harm and Depression Challenges. European Conference on Information Retrieval. Springer, pp 557–563
24. Du J et al (2018) Extracting psychiatric stressors for suicide from social media using deep learning. BMC Med Inform Decis Mak 18(2):43
25. Ma L, Wang Z, Zhang Y (2017) Extracting depression symptoms from social networks and web blogs via text mining. International Symposium on Bioinformatics Research and Applications. Springer, pp 325–330
26. Chen X, Sykora M, Jackson T, Elayan S, Munir F (2018) Tweeting Your Mental Health: an Exploration of Different Classifiers and Features with Emotional Signals in Identifying Mental Health Conditions
27. Wang T, Brede M, Ianni A, Mentzakis E (2017) Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International conference on web search and data mining*. pp. 91–100.
28. De Choudhury M, Counts S, Horvitz E (2013) Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 2013: ACM, pp. 47–56.
29. De Choudhury M, Gamon M, Counts S, Horvitz E (2013) Predicting depression via social media. Icwsm 13:1–10
30. Preoţiuc-Pietro D, Sap M, Schwartz HA, Ungar L (2015) Mental illness detection at the World Well-Being Project for the CLPsych 2015 shared task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 40–45.
31. Burnap P, Colombo W, Scourfield J (2015) Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM conference on hypertext & social media*, 2015, pp. 75–84.
32. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H (2015) Recognizing depression from twitter activity", in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015: ACM, pp. 3187–3196.
33. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M (2014) Affective and content analysis of online depression communities. IEEE Trans Affect Comput 5(3):217–226
34. Yin Z, Sulieman LM, Malin BA (2019) A systematic literature review of machine learning in online personal health data. J Am Med Inform Assoc 26(6):561–576
35. Hu Q, Li A, Heng F, Li J, Zhu T (2015) Predicting depression of social media user on different observation windows. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 2015, vol. 1: IEEE, pp. 361–364.
36. Coppersmith G, Ngo K, Leary R, Wood A (2016) "Exploratory analysis of social media prior to a suicide attempt", in *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, 2016, pp. 106–117.
37. Huang X, Zhang L, Chiu D, Liu T, Li X, Zhu T (2014) Detecting suicidal ideation in Chinese microblogs with psychological lexicons. In *2014 IEEE 11th Intl Conf on Ubiquitous Intelligence and Computing and 2014 IEEE 11th Intl Conf on Autonomic and Trusted Computing and 2014 IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops*, 2014: IEEE, pp. 844–849.
38. Guan L, Hao B, Cheng Q, Yip PS, Zhu T (2015) Identifying Chinese microblog users with high suicide probability using internet-based profile and linguistic features: classification model. JMIR Mental Health. 2(2):e17
39. Saravia E, Chang C-H, De Lorenzo RJ, Chen Y-S (2016) MIDAS: Mental illness detection and analysis via social media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016: IEEE, pp. 1418–1421.
40. Wang Y, Wang Z, Li C, Zhang Y, Wang H (2020) A Multitask Deep Learning Approach for User Depression Detection on Sina Weibo. *arXiv preprint* arXiv:2008.11708

41. Orabi AH, Buddhitha P, Orabi MH, Inkpen D (2018) Deep learning for depression detection of twitter users. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 88–97.
42. Coppersmith G, Harman C, Dredze M (2014) Measuring post traumatic stress disorder in Twitter. In *Eighth international AAAI conference on weblogs and social media*, 2014.
43. Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The development and psychometric properties of LIWC2015
44. Chen X, Sykora MD, Jackson TW, Elayan S (2018) What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pp. 1653–1660.
45. Wilson T *et al*. (2005) OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*
46. Thelwall M, Buckley K, Paltoglou G, Cai D, Kappas A (2010) Sentiment strength detection in short informal text. J Am Soc Inform Sci Technol 61(12):2544–2558
47. Durahim AO, Coşkun M (2015) # iamhappybecause: Gross National Happiness through Twitter analysis and big data. Technol Forecast Soc Chang 99:92–105
48. Bollen J, Gonçalves B, Ruan G, Mao H (2011) Happiness is assortative in online social networks. Artif Life 17(3):237–251
49. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
50. Paul MJ, Dredze M (2011) You are what you tweet: Analyzing twitter for public health. In *Fifth international AAAI conference on weblogs and social media*: Citeseer.
51. Ji S, Yu CP, Fung S-F, Pan S, Long G (2018) Supervised learning for suicidal ideation detection in online user content. Complexity 2018:1–10
52. Kumar A, Garg G (2019) Sentiment analysis of multimodal twitter data. Multimed Tools Appl 78(17):24103–24119
53. Chiu CY, Lane HY, Koh JL, Chen AL (2020) Multimodal depression detection on instagram considering time interval of posts. J Intell Inf Syst 56(1):1–23
54. Reece AG, Danforth CM (2017) Instagram photos reveal predictive markers of depression. EPJ Data Sci 6(1):1–12
55. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, Ungar LH (2019) What twitter profile and posted images reveal about depression and anxiety. In *Proceedings of the International AAAI Conference on Web and Social Media*. 13, 236–246.
56. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
57. Bonta V, Janardhan NKN (2019) A comprehensive study on lexicon based approaches for sentiment analysis. Asian J Comput Sci Technol 8(S2):1–6
58. Association AP (2013) *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.
59. Mohammad SM, Turney PD (2013) Crowdsourcing a word–emotion association lexicon. Comput Intell 29(3):436–465
60. Chua CEH, Storey VC, Li X, Kaul M (2019) Developing insights from social media using semantic lexical chains to mine short text structures. Decis Support Syst 127:113142
61. Dumais ST (2004) Latent semantic analysis. Ann Rev Inf Sci Technol 38(1):188–230
62. Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
63. Mao K, Niu J, Chen H, Wang L, Atiquzzaman M (2018) Mining of marital distress from microblogging social networks: A case study on Sina Weibo. Futur Gener Comput Syst 86:1481–1490
64. 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 Transactions on Computational Social Systems. https://doi.org/10.1109/TCSS.2020.3047604
65. Kim J, Lee J, Park E, Han J (2020) A deep learning model for detecting mental illness from user content onsocial media. Sci Rep 10(1):1–6. https://doi.org/10.1038/s41598-020-68764-y
66. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, Ungar LH (2019) What twitter profile and posted images reveal about depression and anxiety. In: Proceedings of the International AAAI Conference on Web and Social Media, vol 13, pp 236–246
67. Tadesse MM, Lin H, Xu B, Yang L (2019) Detection of depression-related posts in reddit social media forum. IEEE Access 7:44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180

68. Islam MR, Kabir MA, Ahmed A, Kamal AR, Wang H, Ulhaq A (2018) Depression detection from socialnetwork data using machine learning techniques. Health Inf Sci Syst 6(1):1–2. https://doi.org/10.1007/s13755-018-0046-0
69. Ferwerda B, Tkalcic M (2018) You are what you post: What the content of Instagram pictures tells aboutusers' personality. In: The 23rd International on Intelligent User Interfaces, March 7–11, Tokyo, Japan. CEUR-WS
70. Chen X, Sykora M, Jackson T, Elayan S, Munir F. Tweeting your mental health: an exploration of different classifiers and features with emotional signals in identifying mental health conditions