Paper type: Original Paper

Detecting Depression in Users of Online Forums: Enabling Early Healthcare Intervention using Language Models

David Owen[1], MSc; Dimosthenis Antypas[1], MSc; Athanasios Hassoulas[2], PhD; Antonio F. Pardiñas[3], PhD; Luis Espinosa-Anke[1], PhD; Jose Camacho Collados[1], PhD

[1] School of Computer Science and Informatics, Cardiff University, UK
[2] Centre for Medical Education (C4ME), School of Medicine, Cardiff University, UK
[3] MRC Centre for Neuropsychiatric Genetics and Genomics, Division of Psychological Medicine and Clinical Neurosciences, School of Medicine, Cardiff University, UK

Abstract

Background: Major Depressive Disorder (MDD) is a common mental disorder that affects 5% of adults worldwide. Early contact with healthcare services is critical in achieving an accurate diagnosis and improving patient outcomes. Key symptoms of MDD (depression hereafter) such as cognitive distortions are observed in verbal communication, which can manifest in the structure of written language as well. Thus, the automatic analysis of text outputs may provide opportunities for early interventions in settings where written communication is rich and regular, such as social media and online forums.
Objective: The objective was twofold. We sought to gauge the effectiveness of different machine learning approaches to identifying users of the mass online forum Reddit who eventually disclose a diagnosis of depression. We then aimed to determine whether the time between a forum post and a depression diagnosis date is a relevant factor in performing this detection.
Methods: Two Reddit datasets containing posts belonging to users with and without a history of depression diagnosis were obtained. An intersection of these datasets provided users with an estimated date of depression diagnosis. This derived dataset was used as input to several machine learning classifiers, including Transformer-based Language Models.
Results: BERT (Mental Bidirectional Encoder Representations from Transformers) and MentalBERT Transformer-based Language Models proved most effective in distinguishing forum users with a known depression diagnosis from those without. They each obtained a mean F1 score of 0.64 across the experimental setups used for binary classification. The results also suggested that the final 12 to 16 weeks (about 3 to 4 months) of posts prior to a depressed user's estimated diagnosis date are most indicative of their illness, with data prior to that period not helping models detect more accurately. Furthermore, in the four-to-eight-week period prior to the user's estimated diagnosis date, their posts exhibited more negative sentiment than any other four-week period in their post history.

Conclusions: Transformer-based Language Models may be used on data from online social media forums to identify users at risk of psychiatric conditions such as depression. Language features picked up by these classifiers might predate depression onset by weeks to months, enabling proactive mental healthcare interventions to support those at risk of this condition.
Keywords: mental health; depression; internet; natural language processing; transformers; language models; sentiment.

Introduction

Major Depressive Disorder (MDD) is one of the most prevalent mental illnesses worldwide, affecting nearly 5% of adults [1]. Depressive episodes, which are symptoms of MDD and other psychiatric conditions, are even more common with nearly 30% of individuals developing them at least once in their lifetime [2]. Characteristics of MDD and depressive episodes ("depression" hereafter) include low mood, feelings of worthlessness or guilt, and recurrent thoughts of death [3]. Early intervention has been reported to significantly improve patient outcomes and lessen the financial burden on healthcare services [4]. However, the stigma associated with psychiatric conditions such as depression leads to patients underreporting to healthcare services [5,6].

Given that a number of individuals who would normally meet the criteria for depression underreport to healthcare services, consideration should be given to how key symptoms may manifest in written language on social media platforms [7]. Longhand discussion websites such as Reddit are a rich source of such information where users may publish series of posts spanning many months or years [8]. Natural Language Processing (NLP) may be used to identify features in these posts that are predictive of a user who may have depression. Crucially, if affected users are identified prior to formal diagnosis, this may provide an opportunity for early healthcare intervention in these cases.

In this paper we derive a specialized subset of an annotated dataset, which contains Reddit posts belonging to users that have received a diagnosis of depression. The subset allows us to consider posts prior to each user's approximate diagnosis date.

We use state-of-the art and domain specific language models (LMs) to assist in the detection of depression. These LMs outperform baseline approaches in a variety of experimental settings. Notably, they show adeptness in the early detection of depression. Moreover, through our model-analysis we provide an exhaustive analysis on the temporal aspect related to preemptive detection, providing insights on the time depression symptoms materialized prior to the diagnosis. Finally, we investigate the role of sentiment in depressed users' posts and provide a qualitative analysis based on the model performance.

Related Work

There is growing literature on using NLP techniques to analyze depression patterns in social media [9,10].

Yates et al [11] developed an approach for distinguishing forum users who had self-reported a diagnosis of depression from those who did not. It used a Convolutional Neural Network (CNN) to aggregate user posts in a purpose-built dataset, the Reddit Self-reported Depression Diagnosis dataset (RSDD). Their follow-up work involved the conception of a sister dataset, RSDD-Time [12], which contained Reddit posts where users declared a past diagnosis of depression and this diagnosis was linked to an estimated date. Dates were inferred from explicit but often imprecise time expressions in user posts. However, these works did not consider pre-emptive detection of depression amongst the Reddit users in their datasets. That is, they did not consider methods for detecting depression in users prior to their diagnoses.

Recent NLP work has explicitly focused on early depression detection. Pre-emptive detection of mentions of depression amongst Twitter users has been demonstrated with a degree of success by Owen et al [13]. Abed-Esfahani [14] report similar findings using Reddit data. However, both of those works were limited by the uncertainty of whether the users referring to this condition had been formally diagnosed. Shah et al [15] also considered approaches to early detection of depression in Reddit users. In this case it was known whether the users had received a physician's diagnosis. However, it was not certain whether the users' posts occurred before or after their diagnoses because the dates of the diagnoses were not known. To truly gauge the effectiveness of pre-emptive detection methods, a series of user posts prior to a known diagnosis date is required. Eichstaedt et al [16] examined language in Facebook posts that may have been predictive of depression as shown in patients' medical records. They achieved an F1 score of 0.66 via logistic regression modelling that used only the language preceding each patient's depression diagnosis.

This paper therefore also seeks to extend the existing work on pre-emptive depression detection. We will consider social media users whose depression diagnosis date is known and use LMs to harness the language of user posts.

Ren et al [17] performed emotion driven detection of depression on Reddit, achieving F1 scores exceeding 0.9. Their work considered individual depression posts rather than series of posts. Nevertheless, their effective use of emotional semantic information suggested that dissection of our own results could be enhanced using sentiment analysis, which we include in our analysis to provide further insights.

Objectives

We sought to gauge the performance of several machine learning classifiers in the task of distinguishing between RSDD dataset users reporting and not reporting a

diagnosis of depression, which from here onwards we will term as “depressed” and “controls” respectively. We then used the best performing classifier in a temporal driven binary classification task. The purpose was to determine the volume of posts in a depressed user’s post timeline that is most indicative of their illness. To do this we considered only the posts authored prior to depressed users’ estimated diagnosis dates. Moreover, we considered only posts published up to six months prior to those dates.

The motivation for considering this six-month time range hails from Winkour et al [18], and their observation that over 50% of depression patients experienced their first onset at least six months prior to their formal diagnosis. Reece et al [19] made a similar observation when examining Twitter users.

The time during which individuals with symptoms or traits of depression remain undiagnosed poses a serious health risk. Patients who remain undiagnosed thus untreated, experience a worse outcome than would be the case if they were treated [20], particularly after their first episode [21]. Methods for assessing suitable time points for healthcare intervention are needed to identify ways to improve patient outcomes. They are also likely to advance the field of psychiatric therapeutics by supporting modifications to clinical guidelines or the design of randomized-controlled trials [22]. A larger body of evidence on this matter could also help in identifying patients to be targeted for more thorough mental health assessments, and provided with further resources, support, and treatment [23].

Methods

Data Description

Our work is based on the RSDD (Reddit Self-reported Depression Diagnosis) and RSDD-Time datasets [24]. RSDD contains Reddit posts of 9,210 depressed users and 108,731 control users. Posts were published between January 2006 and October 2016. The representation of users in RSDD is depicted in Textbox 1.

Textbox 1. An abstract representation of RSDD user data. It is not permissible to reveal true user IDs, post dates, or post texts due to privacy reasons.

{ user_id: 1, posts: [(<date 1>, <text>), ..., (<date n>, <text>)], label: <either **depressed** or **control**> },
{ user_id: 2, posts: [(<date 1>, <text>), ..., (<date n>, <text>)], label: <either **depressed** or **control**> },
…,
{ user_id: n, posts: [(<date 1>, <text>), ..., (<date n>, <text>)], label: <either **depressed** or **control**> }

RSDD-Time contains 598 annotated Reddit posts, each of which belongs to a user who declares that they have been formally diagnosed with depression. Posts were

published between June 2009 and October 2016. 529 of these posts belong to depressed users that are also present in RSDD.

RSDD-Time annotations include the recency of a user's diagnosis with respect to the date that their post was authored. Permissible recency annotations are as follows:

0 – unspecified, 1 - in the past, 2 - up to 2 months ago, 3 - between 2 months and one year ago, 4 - between 1 year and 3 years ago, and 5 - more than 3 years ago.

The representation of users in RSDD-Time is depicted in Textbox 2.

Textbox 2. An abstract representation of RSDD-Time user data. It is not permissible to reveal true user IDs, diagnosis post texts, or post dates, due to privacy reasons.

{ user_id: 1, diagnosis_post: <text>, post_date: <date>, recency: <**0, 1, 2, 3, 4, or 5**>
},
{ user_id: 2, diagnosis_post: <text>, post_date: <date>, recency: <**0, 1, 2, 3, 4, or 5**>
},
…,
{ user_id: n, diagnosis_post: <text>, post_date: <date>, recency: <**0, 1, 2, 3, 4, or 5**>
}

We used this information to estimate the diagnosis dates of the 529 users present in both RSDD and RSDD-Time. Those with recency annotations of 0 or 1 were ignored since their diagnosis dates could not be estimated with any degree of accuracy. For each of the remaining users, we determined whether the estimated diagnosis date fell between the date of their first RSDD post and the date of their RSDD-Time diagnosis post. 72 depressed users remained.

10 matching control users were then sought for each of the 72 depressed users. To accomplish this, candidate control users were retrieved randomly from RSDD and analyzed sequentially. The candidate's posts dated prior to the corresponding depressed user's estimated diagnosis date were considered. If the number of posts belonging to the candidate did not vary by more than 15% with respect to the depressed user, then the candidate was considered a match. A control user matched in this way was not considered a candidate for subsequent depressed users.

Since sufficient matching control users could not be found for 2 of the depressed users, they were excluded from the resulting dataset. The dataset contained 70 depressed users each of which had 10 matching control users. It therefore totaled 770 users. Posts were published between April 2006 and June 2016. We named our dataset RSDD-Matched. Characteristics of RSDD-Matched are shown in Table 1.

Table 1. Statistics of the RSDD-Matched dataset.

	Depressed	Controls

Total users		
	70	700
Total posts		
	36,826	364,747
Total words		
	1,742,388	8,188,090
Average posts per user		
	526.1	521.1
Average words per post		
	47.3	22.4
Shortest post (words)		
	1	1
Longest post (words)		
	2,642	1,894

Since RSDD does not include posts made in mental health subreddits, a depressed user's diagnosis is certain to not be revealed until the time of their diagnosis post. There is language indicative of mental health conversation in other subreddits, however.

Descriptive analysis of RSDD

To better understand our dataset, we performed a simple descriptive analysis of RSDD. Word-level exploratory analyses on corpora have been extensively used in corpus linguistics and NLP to gain insights on word prominence. Typically, these follow a bag-of-words [25], PMI (Pointwise Mutual Information) [26], or TF-IDF (Term Frequency-Inverse Document Frequency) [27] approach. In our case, we used lexical specificity [28], which is a statistical measure based on the hypergeometric distribution to identify the most prominent words in a corpus. We chose to use lexical specificity because it is structured in a way that is ideal to extract corpus specific vocabulary given a global corpus (RSDD) and its subsets (depressed users and control users) [29]. It is also a more robust metric for term importance when dealing with different lengths of text [30], which is often the case for Reddit posts.

RSDD is partitioned into two subsets, or subcorpora, one containing posts of the depressed users, and another containing posts of the control users. After lemmatizing the corpus, lexical specificity analysis revealed the unigrams (single words) most frequently used by depressed and control users (Table 2). The score column indicates how relevant a unigram appears to be in each subset. For reference, the term "woman" makes up 0.18% of the total words that appear in the depressed users subset compared to only 0.06% of the control users subset.

Table 2. Top ranked words of RSDD depressed and control users in terms of lexical specificity.

	Word	Score
Depressed users		
	people	338,131.45
	know	164,368.51
	thing	150,440.49
	feel	118,483.23
	time	97,250.09
	woman	96,165.35
	go	79,611.79
	want	75,379.17
	life	67,769.01
	relationship	62,606.64
Control users		
	game	42,234.94
	trade	39,445.65
	key	30,031.17
	team	24,333.73
	play	17,389.38
	player	16,186.61
	shiny	14,032.27
	hatch	13,265.87
	thank	10,177.49
	add	10,005.14

To put the results into context, we should mention that a lexical specificity score of X for a given word W with frequency F means that the probability of W occurring at least F times in the subcorpus is lower than 10^-X (assuming a random distribution). For instance, a lexical specificity score of 42,234 for "game" means that the probability of "game" having a frequency of f= 5,373,938 or higher in the control users subcorpus is 10^- 42,234 (i.e. - an exceptionally low probability which means "game" is overrepresented in the control users subset). In general, we can observe a pattern in which depressed users tend to employ more relationship or family-related words (e.g. - "woman", "relationship") and words related to the depression symptoms themselves (e.g. - "life"). In contrast, control users seem to employ more mundane terms related to the Subreddit communities, such as game-related terms (e.g. - "game" or "team"). While this analysis is only based on the statistical frequency of the terms employed, it may provide further evidence that developing automatic methods to identify users with depression may indeed be feasible. In the Error Analysis sections we extend this initial inspection to better understand the errors made by automatic models.

Methodology

In this section, we provide more details of our proposed methods to tackle the depression detection task. Framing the task as a machine learning problem, we

considered nine methods based on linear classifiers and more recent language models.

The initial baselines entailed an SVM (Support Vector Machine) architecture. An SVM is an algorithm that learns by example to assign labels to objects [31]. In our case, the objects are Reddit users, and permissible labels are "depressed" and "control." SVMs have demonstrated effectiveness in the detection of depression-related posts in Reddit [8,32]. Our SVM configurations used different features derived from the user posts. These features included TF-IDF, word embeddings, and finally, a combination of both TF-IDF and word embeddings. TF-IDF [33] features represent the words deemed most significant amongst the user posts. A word embedding is a real-valued vector representation of a word [34]. Words of similar meaning have vectors of similar value.

The SVM model used was that of scikit-learn [35], as was the TF-IDF vectorizer implementation. The word embeddings generated for each Reddit post were drawn from GloVe vectors (Global Vectors) trained on Wikipedia and Gigaword data [36]. These vectors had a dimensionality of 300, and so did the averaged embedding generated. We performed Reddit post text preprocessing prior to their input to the SVM. All posts underwent quotation normalization, so each quotation character was represented by a single apostrophe. All new line and carriage return characters were replaced with spaces so that posts were represented as a single line string. The posts were then concatenated on a per-user basis so that each user's posting history was represented as one single line string. The SVM used a linear kernel, which is appropriate for text classification problems [37,38,39].

The remaining six classifiers we considered were transformer based LMs. LMs are a statistical means of predicting words [40], while transformers provide a neural network-based approach to generating such models [41]. Transformer-based LMs have proved effective in detecting psychiatric illness related Reddit posts [12,42,43]. We therefore chose to use transformer based LMs to support the detection of depression in RSDD-Matched. We chose BERT [44] and ALBERT [45], which are appropriate for a wide variety of application areas. We also chose four specialist language models BioBERT [46], Longformer [47], MentalBERT [48], and MentalRoBERTa [48]. BioBERT is apt for use where biomedical concepts are prevalent, such as electronic medical records [49], patient descriptions [50], and health-related Twitter postings [51]. Longformer is designed for use where text is formed of long documents. Indeed, there are posts in RSDD-Matched that exceed 2,000 words. Finally, MentalBERT and MentalRoBERTa are customized for the domain of mental healthcare, having been trained using text drawn from mental health discussion forums.

All six transformer-based LMs are pre-trained bidirectional language representations. This means that for any given word in a text segment, its neighboring words to both the left and right are examined so that the context of the word is well understood. These representations lend themselves to high

performance in text classification tasks when compared with traditional approaches using SVMs, for example [52,53].

We used the Simple Transformers software library [54] to deploy the LMs. The library provides an Application Programming Interface (API) to the Transformers Library, which itself provides access to BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa models [55]. The BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa classifiers used were "bert-base-uncased", "albert-base-v1", "biobert-base-cased-v1.1", "longformer-base-4096", "mental-bert-base-uncased", and "mental-roberta-base" respectively. In addition to Simple Transformers' default hyperparameters, the LM classifiers were instantiated with sliding window enabled. Transformer-based LMs may consume only a limited number of tokens (akin to words), typically 512. Since the posting histories of most users in RSDD-Matched exceed 512 words, a specialist approach to applying LMs to these posts is needed. Sliding window is one such approach [56].

Experimental Setup

Preemptive Depression Identification Experiment

This first experiment consisted of examining the performance of several machine learning classifiers in the task of distinguishing between depressed and control users in RSDD-Matched. The purpose of this experiment is to understand to what extent preemptive detection of depression in social media is possible. Moreover, this experiment is aimed at understanding the capabilities of machine learning classifiers for this task, and the suitability of different methods in the task. The results were used to provide a competitive model for the subsequent finer-grained temporal experiment.

We used nine different classifiers. Three entailed an SVM as described in the Methodology section. The remaining six were the BERT, ALBERT, BioBERT, Longformer, MentalBERT, and MentalRoBERTa, which are also described in the Methodology section.

In addition to the above classifiers, we included a naïve baseline that predicts positive instances in all cases.

Since the number of positive instances (i.e. - depressed users) in RSDD-Matched is small, we chose not to use a traditional train-test split. Instead, we used 5-fold cross-validation; an approach also used by Eichstaedt et al [14]. Further, we varied the number of matching control users across four iterations of the experiment (Table 3).

Table 3. Variations of the preemptive depression identification experiment in terms of the number of matching control users considered.

	Depressed users	Matching control users per	Total users

		depressed user	
Variation 1			
	70	1	140
Variation 2			
	70	3	280
Variation 3			
	70	5	420
Variation 4			
	70	10	770

The purpose of these variations is to test the performance of the classifiers versus increasingly imbalanced datasets. This mimics the conditions likely to be observed in online forums where the number of positive instances (i.e. - depressed users) is dwarfed by the number of negative instances (i.e. - non-depressed users).

Temporal Experiment

The purpose of the second primary experiment was to determine which posting period in a depressed user's post timeline is most indicative of depression. This involved using a subset of RSDD-Matched. The performance of binary classifiers versus temporal subsets of the posts in the six months prior to the users' estimated diagnosis dates were measured.

The RSDD-Matched subset contained only depressed users who had at least one post in the two weeks prior to their estimated diagnosis date. Of the 70 depressed users in our RSDD subset, 14 of them did not have any posts in this two-week period. As a result, we used only 56 depressed users in the temporal experiment. Furthermore, not all the 10 control users matched with each of the 56 depressed were useable since some did not have at least one post in this two-week period. Thus, we performed additional random exclusions of controls to re-balance the dataset. After these exclusions, the dataset used in the temporal experiment contained 56 depressed users each of which had 3 matching control users, totaling 224 users.

The results of the preemptive depression identification experiment were used to partially inform the design of the temporal experiment. Since BERT scored the highest average F1 score across all runs of the preemptive depression identification experiment, it was decided that this be the sole general-purpose transformer-based LM to be used in the temporal experiment. Likewise, MentalBERT scored the highest average F1 score, so was selected as the sole specialist LM. The three variations of the SVM classifier used in the preemptive depression identification experiment were used once again.

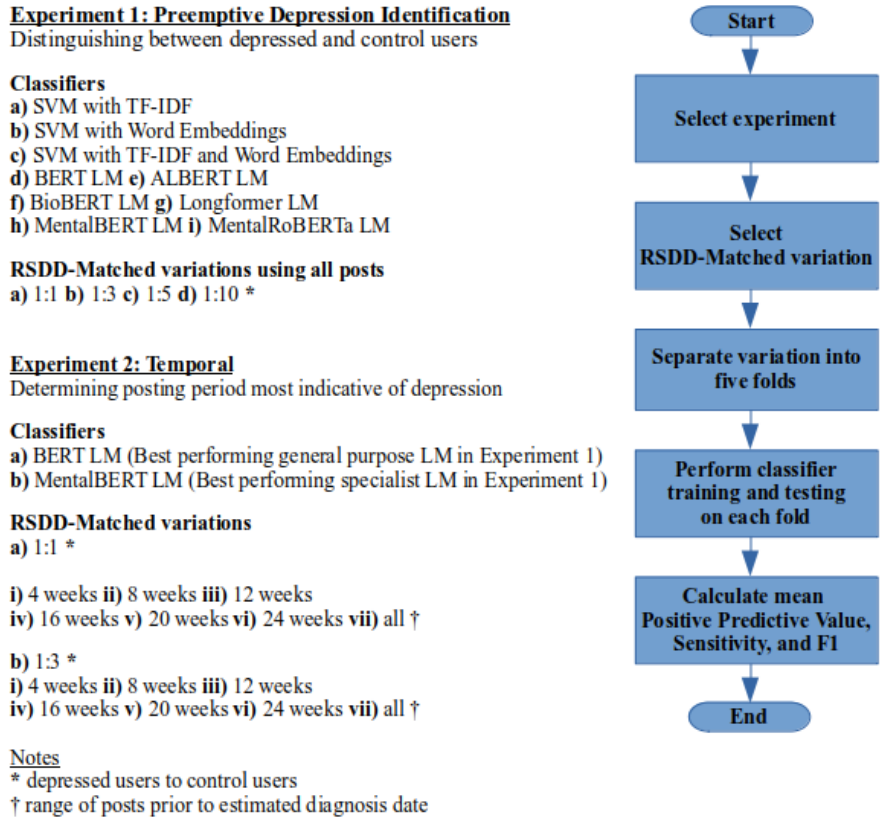
Once again, we used 5-fold cross-validation. Two chief variations of the RSDD-Matched subset and several different temporal configurations were used (Table 4).

Table 4. Variations of the temporal experiment in terms of the number of matching control users and numbers of weeks of posts prior to estimated diagnosis dates considered.

	Depressed users	Matching control users per depressed user	Total users	Weeks of posts omitted prior to estimated diagnosis date	Weeks of posts included prior to estimated diagnosis date
Variation 1					
	56	1	112	0, 4, 8, 12, 16, 20	4, 8, 12, 16, 20, 24
Variation 2					
	56	3	224	0, 4, 8, 12, 16, 20	4, 8, 12, 16, 20, 24

The two chief strands to our experimental setup are summarized in Figure 1.

Figure 1. Summary of the two chief experimental setups.



Sentiment Analysis

We complemented the temporal experiment with a sentiment analysis. The purpose was to identify whether there is a link between sentiment and depression with respect to user posts. The sentiment of a text has been extensively utilized as a predictor to detect signs of depressive moods in microblog users [57,58,59]. Specifically, negatively charged text has often been correlated with depression via expressions of low mood and suicidal ideation [60]. Approaches used to extract sentiment from social media posts have included use of LMs [61] and lexicons such as VADER [62].

To determine whether there is a relation between sentiment and depression, we utilized BERTweet-sentiment, a state-of-the-art transformer model, to classify each post in RSDD-Matched as either negative, neutral, or positive. BERTweet-sentiment is based on the BERTweet [63] implementation, which is trained on a large Twitter corpus and is finetuned for sentiment analysis. Although the model is not trained on Reddit data, we believe that there are enough overlapping lexical characteristics between the two domains in terms of internet slang and text lengths that justify its use.

Our sentiment analysis focused on changes in the sentiment distribution of depressed and control users through time. In step with the design of our temporal experiment, each user's posts are divided into six temporal bands, namely 0 to 3, 4 to 7, 8 to 11, 12 to 15, 16 to 19, and 20 to 23 weeks prior to their estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user). The average percentage of each sentiment in each band is considered.

To establish whether the diagnosis is associated with the sentiment of a post, two regression models were used. The first was based on the lme4 framework [64] and the second used mgcv [65]. Implementations used were those of the R v4.02 statistical environment [66]. We set our outcome variable to be whether a post is "sentimental" (that is, either negative or positive) or not (neutral) and a logistic mixed-effects regression is fitted using all the available posts with the individual user identifier as random effect term. As fixed effects, we used the estimated depression diagnosis (that is, either depressed or control), the time to estimated diagnosis in weeks, the post's word count, and the interaction term of estimated diagnosis with time.

Having sought to establish whether the diagnosis of the user is associated with the sentimentality inferred for each post, we also considered a more fine-grained Multinomial regression model. This is equivalent to fitting a series of logistic models against a reference category [67] and similar to the "stacked" designs used in other

disciplines [68]. For our purposes, we will consider "neutral" as the reference category of our multinomial outcome, so all effect sizes will indicate the probability of a post being positive or negative *instead of* neutral.

Results

Preemptive Depression Identification Experiment

Results of the preemptive depression identification experiment are presented below (Tables 5, 6, 7, and 8). Each table portrays a variation in the number of matched control users used. Positive Predictive Value, Sensitivity, and F1 score are used to measure performance in each variation. Positive Predictive Value denotes how many users classified as depressed, were indeed depressed. Sensitivity denotes how many of the depressed users were correctly classified as depressed. F1 score, which is the harmonic mean of positive predictive value and sensitivity, is apt for use with datasets like ours where class distribution (of depressed and controls) is uneven [69]. By contrast, Accuracy is not apt for use with such datasets [70]. We therefore use F1 as our primary performance metric.

Using F1 as a primary performance indicator, MentalBERT performs best across the variations.

Table 5. Binary classification scores using all posts of 70 depressed users and 1 of their matched control users. Language Model experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

	Positive Predictive Value		Sensitivity		F1	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
SVM using TF-IDF						
	0.637	N/A	0.557	N/A	0.590	N/A
SVM using Word embeddings						
	0.558	N/A	0.543	N/A	0.548	N/A
SVM using TF-IDF and Word embeddings						
	0.673	N/A	0.557	N/A	0.596	N/A
BERT LM						
	0.638	0.021	0.805	0.022	0.709	0.012
ALBERT LM						
	0.606	0.008	0.786	0.015	0.683	0.010
BioBERT LM						
	0.601	0.005	0.862	0.022	0.707	0.005
Longformer LM						

	0.633	0.009	0.838	0.036	0.719	0.018
MentalBERT LM						
	0.660	0.019	0.848	0.008	0.738	0.013
MentalRoBERTa LM						
	0.629	0.002	0.819	0.022	0.709	0.006
Naïve baseline – all depression						
	0.500	N/A	1.000	N/A	0.667	N/A

Table 6. Binary classification scores using all posts of 70 depressed users and 3 of their matched control users. Language Model experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

	Positive Predictive Value		Sensitivity		F1	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
SVM using TF-IDF						
	0.800	N/A	0.086	N/A	0.153	N/A
SVM using Word embeddings						
	0.411	N/A	0.529	N/A	0.459	N/A
SVM using TF-IDF and Word embeddings						
	0.800	N/A	0.057	N/A	0.107	N/A
BERT LM						
	0.653	0.033	0.481	0.022	0.546	0.025
ALBERT LM						
	0.652	0.034	0.476	0.009	0.547	0.018
BioBERT LM						
	0.654	0.028	0.410	0.030	0.496	0.020
Longformer LM						
	0.653	0.036	0.476	0.036	0.534	0.031
MentalBERT LM						
	0.657	0.034	0.509	0.008	0.562	0.016
MentalRoBERTa LM						
	0.614	0.023	0.471	0.015	0.522	0.002
Naïve baseline – all depression						
	0.250	N/A	1.000	N/A	0.167	N/A

Table 7. Binary classification scores using all posts of 70 depressed users and 5 of their matched control users. Language Model experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

	Positive Predictive Value		Sensitivity		F1	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
SVM using TF-IDF						
	0.400	N/A	0.029	N/A	0.053	N/A
SVM using Word embeddings						
	0.309	N/A	0.471	N/A	0.372	N/A
SVM using TF-IDF and Word embeddings						
	0.200	N/A	0.014	N/A	0.027	N/A
BERT LM						
	0.615	0.028	0.290	0.022	0.379	0.017
ALBERT LM						
	0.555	0.030	0.281	0.009	0.354	0.006
BioBERT LM						
	0.627	0.034	0.252	0.021	0.331	0.027
Longformer LM						
	0.624	0.108	0.286	0.038	0.363	0.059
MentalBERT LM						
	0.572	0.002	0.329	0.043	0.400	0.040
MentalRoBERTa LM						
	0.562	0.027	0.343	0.000	0.419	0.010
Naïve baseline – all depression						
	0.167	N/A	1.000	N/A	0.286	N/A

Table 8. Binary classification scores using all posts of 70 depressed users and 10 of their matched control users. Language Model experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

	Positive Predictive Value		Sensitivity		F1	
	Mean	St Dev	Mean	St Dev	Mean	St Dev
SVM using TF-IDF						
	0.000	N/A	0.000	N/A	0.000	N/A
SVM using Word embeddings						

	0.212	N/A	0.371	N/A	0.268	N/A
SVM using TF-IDF and Word embeddings						
	0.000	N/A	0.000	N/A	0.000	N/A
BERT LM						
	0.100	0.000	0.014	0.000	0.025	0.00
ALBERT LM						
	0.089	0.019	0.014	0.000	0.025	0.001
BioBERT LM						
	0.067	0.115	0.005	0.008	0.009	0.016
Longformer LM						
	0.024	0.019	0.019	0.033	0.021	0.037
MentalBERT LM						
	0.167	0.058	0.014	0.000	0.026	0.001
MentalRoBERTa LM						
	0.272	0.185	0.034	0.008	0.057	0.018
Naïve baseline – all depression						
	0.091	N/A	1.000	N/A	0.167	N/A

Word embeddings (vector representations) make for strong sensitivity (recall), while TF-IDF features cause deficient performance. Positive predictive value (precision) is observed best when using the specialist LM, MentalBERT. Best F1 is also achieved by MentalBERT and exceeds the naïve baseline.

Error Analysis

We now consider selected users from RSDD-Matched and the performance of the classifiers against them. We will examine one misclassified user per variation of the experiment (in terms of depressed users and the number of matched controls). For each variation, we will examine the strongest performing classifier and the user that it misclassified with highest probability.

To identify potential reasons for the misclassifications, we examine lexical properties of the user posts using three approaches. The first approach involves ascertaining the chief topic conveyed by the posts - a topic being represented by five words. Topic modelling via LDA (Latent Dirichlet Allocation) is used to accomplish this [71,72]. The second approach examines the chief TF-IDF features of the user posts. The third approach is to count the frequencies of the depressed and control vocabularies (Table 2) that appear across the posts.

We present misclassified depressed users with respect to each variation of the experiment (Table 9). We also present misclassified control users with respect to each variation (Table 10).

One depressed user is widely misclassified. User d13 is deemed a control user by three different classifiers across three different variations. While depressed vocabulary counts slightly outweigh their control counterparts, the totals for both vocabularies are nominal. The topic of the user's posts is probably more indicative of reasons for misclassification. Certainly, a theme concerning death or dying appears to be present, but this is diluted by optimistic sounding references of a temporal and geographical nature. Further diluting references are revealed among the TF-IDF features where strong terms such as "love" are present. It would seem that the classifiers construe such references as those belonging to a control user.

User d38 is perhaps misclassified for similar reasons. Counts for both depressed and control vocabularies are small. Positive terms such as "welcome" and "invite" might be deemed to belong to a control user.

Inferior performance is seen across the classifiers in the most imbalanced environment. We examine user depressed user d57, which has been misclassified with a probability close to certainty. The depressed vocabulary count dwarfs the control vocabulary count. However, the classifier seems to harness the overarching nature of the user's posts, as indicated by the topic model and TF-IDF features, when making its decision. A prevalence of "good" natured posts will inevitably see the user deemed similar to a control user when represented in vector space.

Table 9. Depressed users most strongly misclassified in each variation of the preemptive depression identification experiment. Lexical properties of those users' posts are provided.

		Selected depressed users misclassified as control users			
Depressed : Controls					
		1:1	1:3	1:5	1:10
Classifier					
		MentalBERT LM	MentalBERT LM	MentalRoBERTa LM	SVM using Word embeddings
User					
		d13	d38	d13	d57
Control probability					
		0.93	0.94	0.99	0.98
Sum of post					

lengths in words					
		1,696	1,888	1,696	55,897
Topic					
		news hawaii time dead blue	sir-geo welcomed invite leave warlock	news hawaii time dead blue	good time people years problem
Chief TF-IDF features					
		love minnesota diablo time man bud zoidberg like month hawaii	sir geo welcome invite warlock leave titan psn run need	love minnesota diablo time man bud zoidberg like month hawaii	good know use make time thank link want try like
Depressed vocab counts					
	people	1	1	1	64
	know	6	0	6	93
	thing	3	0	3	35
	feel	2	2	2	10
	time	5	8	5	99
	woman	1	0	1	7
	go	3	0	3	54
	want	3	1	3	71
	life	2	0	2	28
	relationship	0	0	0	2
Control vocab counts					
	game	0	1	0	9
	trade	0	0	0	2
	key	0	0	0	4
	team	2	3	2	4
	play	0	1	0	35
	player	0	0	0	8

	shiny	0	0	0	0
	hatch	0	0	0	0
	thank	1	1	0	15
	add	0	2	0	14

We now consider the misclassified control users with respect to each variation of the experiment (Table 10).

Certain users appear confounding across a number of different classifiers and variations. User c13 is strongly misclassified as a depressed user by both MentalBERT and MentalRoBERTa in the relatively noisy environments of three and five matched control users respectively (Table 10). Depressed vocabulary counts far outweigh control vocabulary counts for this user. Also, the theological topic and TF-IDF features of the user's posts are deemed likely to be that of a depressed user, according to the classifier.

MentalBERT demonstrates adeptness in the most balanced variation of the experiment. We seek possible explanations for its misclassification of user c521. The control vocabulary count slightly outweighs that of the depressed vocabulary. What is more, the topic model and TF-IDF features are composed of terms that tend to complement the control vocabulary. Intuitive reasons for the misclassification as a depressed user are difficult to cite. Therefore, it is possible that in the balanced environment, the classifier simply has too few control users to compare and contrast with the depressed users.

In the noisiest environment, the simpler, word-based model (SVM using Word Embeddings) demonstrates the strongest performance. The transformer-based language models are barely able to perform. The vocabulary of the most strongly misclassified user in this case (c535) offers only a tenuous explanation. The count of depressed vocabulary is small, although it outweighs that of the control vocabulary. However, the topic and TF-IDF terms appear to complement the depressed vocabulary, which may have been a cause of the misclassification.

Table 10. Control users most strongly misclassified in each variation of the preemptive depression identification experiment. Lexical properties of those users' posts are provided.

		Selected control users misclassified as depressed users			
Depressed : Controls					
		1:1	1:3	1:5	1:10
Classifier					
		MentalBER	MentalBER	MentalRoBERT	SVM using

		T LM	T LM	a LM	Word embeddings
User					
		c521	c13	c13	c535
Depressed probability					
		0.99	0.95	0.91	0.91
Sum of post lengths in words					
		1,513	8,489	8,489	1,595
Topic					
		elo play team bronze games	god jesus people good life	god jesus people good life	people shit reddit guy man
Chief TF-IDF features					
		team just suck elo play game like good sydtko win	god think way thing try know jesus people say like	god think way thing try know jesus people say like	say thank guy people reddit man make tell watch let
Depressed vocab counts					
	people	4	48	48	6
	know	2	36	36	3
	thing	3	28	28	1
	feel	1	6	6	1
	time	2	6	6	4
	woman	0	4	4	0
	go	0	4	4	5
	want	3	16	16	1
	life	0	46	46	1

	relationship	0	8	8	0
Control vocab counts					
	game	7	0	0	0
	trade	0	0	0	0
	key	0	0	0	0
	team	9	0	0	0
	play	9	6	6	0
	player	2	0	0	0
	shiny	0	0	0	0
	hatch	0	0	0	0
	thank	1	4	4	1
	add	1	0	0	0

Temporal Experiment

We then proceeded to perform the temporal experiment. Since BERT achieved the highest F1 score across all preemptive depression identification experiment variations, it was selected as the exclusive general-purpose LM here. For the same reason, MentalBERT was selected as the exclusive specialist LM. Results are shown below (Tables 11 and 12). Each table portrays a variation in the number of matched control users used. The average performance of each LM across the two variations is illustrated in Figure 2.

For BERT, strongest sensitivity and F1 scores were seen when only 12 weeks (about 3 months) of posts prior to the estimated diagnosis dates were considered. Subsets larger or smaller than 12 weeks caused a degradation of classifier performance. For MentalBERT, strongest sensitivity and F1 scores were when either 16 or 24 weeks of posts were considered. With BERT scoring a higher F1 at 12 weeks than MentalBERT it suggests that the final 12 weeks of posts prior to a depressed user's estimated diagnosis date may be the most indicative of their illness.

An explanation for the slightly inferior performance of MentalBERT may be found in its construction - it is pretrained on text from mental health subreddits such as "r/depression" and "r/mentalhealth" [48]. RSDD (from which we derived RSDD-Matched), however, does not contain posts from mental health subreddits. Therefore, when RSDD-Matched data is limited as it is in our temporal experiment, more general-purpose models such as BERT may be able to realize stronger performance. BERT is pretrained on more general corpora such as Wikipedia [44].

Table 11. Binary classification scores using 56 depressed users and 1 of their matched control users and six temporal post subsets. The classifiers used are BERT

LM and MentalBERT LM, both of whose experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

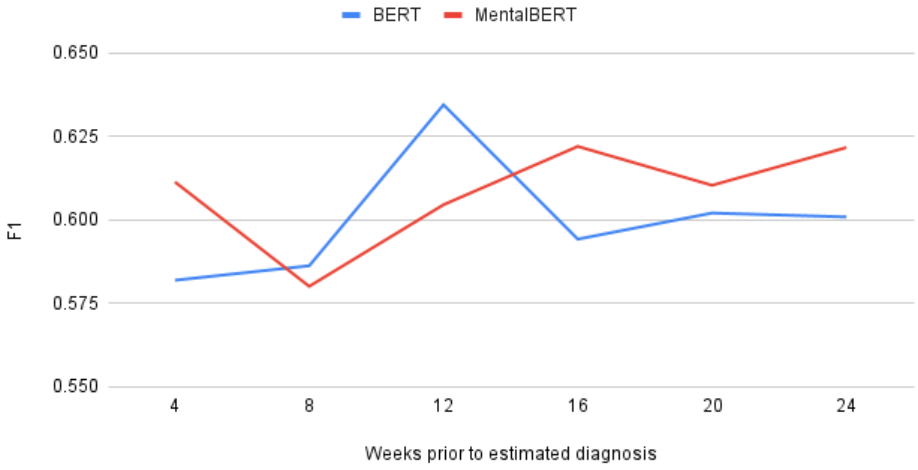
		Positive Predictive Value		Sensitivity		F1	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Last 4 weeks							
	BERT LM	0.575	0.027	0.830	0.039	0.675	0.023
	MentalBERT LM	0.612	0.026	0.835	0.026	0.698	0.017
Last 8 weeks							
	BERT LM	0.598	0.026	0.854	0.071	0.700	0.037
	MentalBERT LM	0.603	0.020	0.842	0.047	0.699	0.022
Last 12 weeks							
	BERT LM	0.605	0.014	0.912	0.018	0.726	0.015
	MentalBERT LM	0.600	0.013	0.888	0.010	0.715	0.008
Last 16 weeks							
	BERT LM	0.570	0.009	0.863	0.026	0.684	0.007
	MentalBERT LM	0.575	0.009	0.907	0.028	0.703	0.016
Last 20 weeks							
	BERT LM	0.569	0.023	0.893	0.036	0.694	0.025
	MentalBERT LM	0.578	0.018	0.882	0.027	0.696	0.014
Last 24 weeks							
	BERT LM	0.565	0.021	0.871	0.027	0.683	0.010
	MentalBERT LM	0.591	0.014	0.890	0.010	0.707	0.011
All posts							
	BERT LM	0.627	0.018	0.824	0.032	0.710	0.019
	MentalBERT LM	0.638	0.009	0.861	0.000	0.732	0.006
Naïve baseline							
		0.500	N/A	1.000	N/A	0.667	N/A

Table 12. Binary classification scores using 56 depressed users and 3 of their matched control users and six temporal post subsets. The classifiers used are BERT LM and MentalBERT LM, both of whose experiments were run three times each, therefore both mean and mean standard deviation scores are provided.

		Positive Predictive Value		Sensitivity		F1	
		Mean	St Dev	Mean	St Dev	Mean	St Dev
Last 4 weeks							
	BERT LM	0.480	0.027	0.538	0.019	0.489	0.010
	MentalBERT LM	0.494	0.019	0.577	0.009	0.525	0.007
Last 8 weeks							
	BERT LM	0.446	0.032	0.538	0.036	0.472	0.035
	MentalBERT LM	0.427	0.027	0.524	0.029	0.461	0.023
Last 12 weeks							
	BERT LM	0.498	0.031	0.619	0.037	0.543	0.035
	MentalBERT LM	0.448	0.007	0.569	0.017	0.494	0.009
Last 16 weeks							
	BERT LM	0.471	0.010	0.565	0.021	0.504	0.011
	MentalBERT LM	0.481	0.023	0.643	0.037	0.541	0.028
Last 20 weeks							
	BERT LM	0.475	0.039	0.577	0.037	0.510	0.034
	MentalBERT LM	0.487	0.018	0.595	0.011	0.524	0.009
Last 24 weeks							
	BERT LM	0.470	0.033	0.591	0.036	0.518	0.033
	MentalBERT LM	0.501	0.022	0.591	0.018	0.536	0.022
All posts							
	BERT LM	0.625	0.021	0.519	0.032	0.562	0.015
	MentalBERT LM	0.588	0.005	0.508	0.010	0.540	0.003

Naïve baseline							
		0.250	N/A	1.000	N/A	0.400	N/A

Figure 2. Average performances of BERT and MentalBERT between 4 and 24 weeks prior to the estimated diagnosis date.



Error Analysis

We once again consider selected users from RSDD-Matched and the performance of the classifiers against them. We again examine one misclassified user per variation of the experiment (in terms of depressed users and the number of matched controls). For each variation, we will examine the strongest performing time span and the user misclassified with highest probability. To identify reasons for the misclassifications, we again examine lexical properties of the user posts using topic models, TF-IDF features, and vocabulary (Table 2) frequency counts.

Misclassified depressed users with respect to the two variations of the experiment are presented below (Table 13).

User d52 is a depressed user misclassified in both balanced and imbalanced environments where only the final 12 weeks of their posts are considered. The vocabulary of these posts intersects with very little of the chief depressed vocabulary. It intersects with slightly more of the chief control vocabulary. The topic and TF-IDF features, intuitively speaking, appear to belong to that of a control rather than a depressed user. Perhaps the balanced environment with temporally limited post histories provides little training data from which the classifier can learn to differentiate between controls and depressed users. These cases, while rare, may occur in practice, and highlight the importance of being careful in over-relying on automatic models for individual assessments without human expert intervention.

Table 13. Depressed users most strongly misclassified in each variation of the temporal experiment. Lexical properties of those users' posts are provided.

		Selected depressed users misclassified as control users	
Depressed: Controls			
		1:1	1:3
Time span			
		Last 12 weeks	Last 12 weeks
Classifier			
		BERT LM	BERT LM
User			
		d52	d52
Control probability			
		0.869	0.935
Sum of post lengths in words			
		1,225	1,225
Topic			
		england belgium hamster time team	england belgium hamster time team
Chief TF-IDF features			
		thank team player help time goal cage post second start	thank team player help time goal cage post second start
Depressed vocab counts			
	people	0	0
	know	1	1
	thing	1	1
	feel	0	0

	time	4	4
	woman	0	0
	go	0	0
	want	2	2
	life	0	0
	relationship	0	0
Control vocab counts			
	game	2	2
	trade	0	0
	key	0	0
	team	4	4
	play	0	0
	player	1	1
	shiny	0	0
	hatch	0	0
	thank	2	2
	add	1	1

We now consider the misclassified control users with respect to the two variations of the experiment (Table 14).

We first consider user c481. Both its depressed and control vocabulary counts are zero, which offers some insight into the misclassification. The topic and TF-IDF features of the posts appear to align with that of a control user. However, it is likely that prevalence of "pain" is a confounding factor. This term might be intuitively linked with depressed users, so it may mislead the classifier. And again, the limited temporal range of posts in this setting provides little data from which the classifier can learn.

User c13 is a confounder in the preemptive depression identification experiment and has proven to be so in the temporal experiment. Even when considering only the last 12 weeks of the user's posts in an imbalanced environment, the theologically themed vocabulary is not diluted. It intersects strongly with the vocabulary of the depressed users and provides an explanation for this misclassification.

Table 14. Control users most strongly misclassified in each variation of the temporal experiment. Lexical properties of those users' posts are provided.

		Selected control users misclassified as depressed users	
Depressed: Controls			
		1:1	1:3

Time span			
		Last 12 weeks	Last 12 weeks
Classifier			
		BERT LM	BERT LM
User			
		c481	c13
Depressed probability			
		0.963	0.917
Total length of posts in words			
		258	8,489
Topic			
		food clove tomorrow pain suspect	god jesus people good life
Chief TF-IDF features			
		reply eat food cat clove pain suspect tooth vet water	god think way thing try know jesus people say like
Depressed vocab counts			
	people	0	24
	know	0	18
	thing	0	14
	feel	0	3
	time	0	3
	woman	0	2
	go	0	2
	want	0	8
	life	0	23
	relationship	0	4
Control			

vocab counts			
	game	0	0
	trade	0	0
	key	0	0
	team	0	0
	play	0	3
	player	0	0
	shiny	0	0
	hatch	0	0
	thank	0	2
	add	0	0

Sentiment Analysis

The sentiment analysis to complement the temporal experiment was then performed. We present the band-wise changes in sentiment for each class (Figures 3 and 4). It is observable that negatively charged posts for depressed users are less frequent as we approach the (estimated) diagnosis date, which may be deemed counter intuitive (Figure 3). However, it is also notable that the depressed users' posts are on average more negative than those of control users throughout the 24-week period (Figure 4). This aligns with previous studies that find a positive correlation between mental illness and negative sentiment [73].

Figure 3: Change in the average percentage of positive and negative posts across six temporal bands: 0 to 3, 4 to 7, 8 to 11, 12 to 15, 16 to 19, and 20 to 23 weeks prior to the estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user).

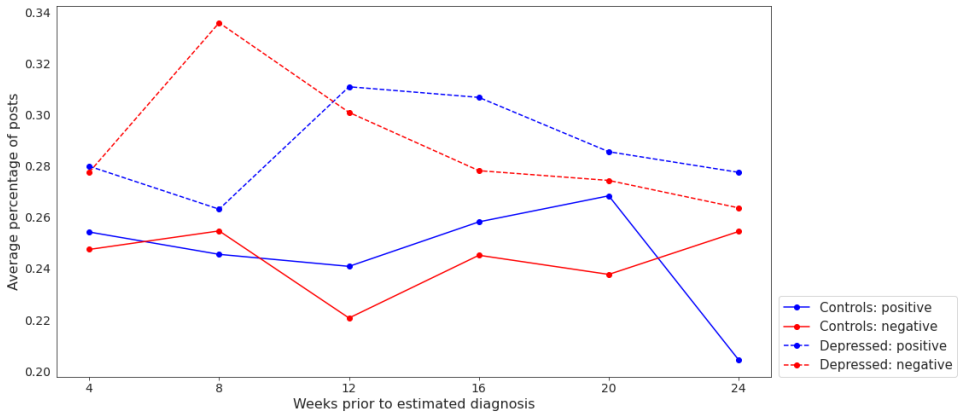
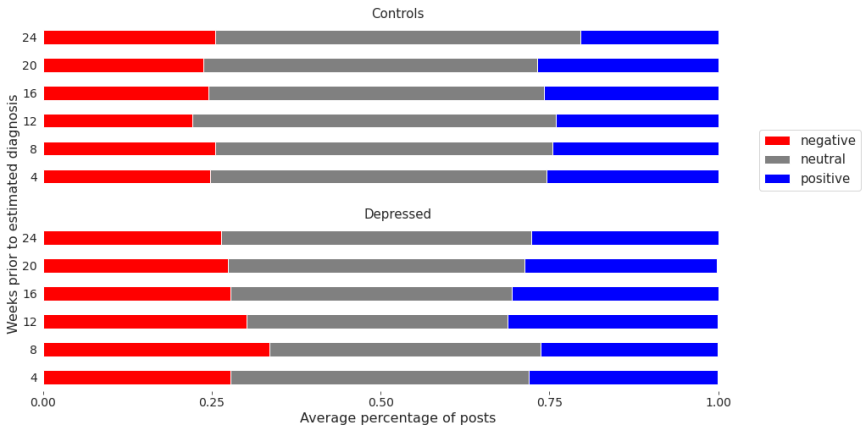


Figure 4: Average percentage of positive and negative posts per temporal band. Temporal bands include 0 to 3, 4 to 7, 8 to 11, 12 to 15, 16 to 19, and 20 to 23 weeks prior to the estimated diagnosis date (for a control user, this is the estimated diagnosis of its matched depressed user).



We then looked to establish whether the diagnosis is associated with the sentiment of a post. The results of the logistic regression model (Table 15) indicate that there is a clear significant association between the diagnosis and the "sentimentality" of the post ($p < 0.05$), despite no apparent effect of temporality. Interestingly, the word count of a post appeared as a significant covariate of this model ($p=0.001$), indicating that longer posts are slightly more likely to be classified as "sentimental", irrespective of the depression status of the user.

Table 15: Logistic regression results for predicting whether a post is neutral or not neutral.

Variable	**Beta**	**Odds-Ratio**	**SE**	**p-value**
Depression diagnosis				
	0.163	1.177	0.035	3.66e-06
Time to diagnosis				
	-0.004	0.996	0.013	0.750
Post word count				
	0.040	1.041	0.012	0.001
Interaction (diagnosis * time)				
	0.011	1.011	0.013	0.406

Table 16: Multinomial regression results for predicting whether a post is positive or negative.

Sentiment	**Variable**	**Beta**	**Odds-Ratio**	**SE**	**p-value**
Positive	Depression				

	diagnosis				
		0.190	1.209	0.047	5.09e-05
	Time to diagnosis				
		0.015	1.015	0.016	0.365
	Post word count				
		-0.070	0.932	0.019	2.21e-04
	Interaction (diagnosis * time)				
		0.045	1.046	0.016	0.006
Negative	Depression diagnosis				
		0.151	1.163	0.041	2.45e-04
	Time to diagnosis				
		-0.019	0.981	0.016	0.243
	Post word count				
		0.103	1.108	0.014	1.20e-12
	Interaction (diagnosis * time)				
		-0.021	0.979	0.016	0.184

Results of the Multinomial Regression Model are presented above (Table 16). Again, all effect size estimates are compatible with our inferences based on the simpler logistic model. However, the multinomial analysis gives us an additional perspective: the effects of the depression diagnosis are similar between positive and negative sentiments, with overlapping confidence intervals statistically indistinguishable. This is the case despite the varying effects of other covariates, such as the word count which displays regression beta coefficients of opposite signs in both sentiments (more words associate with negative posts while fewer words associate with positive posts).

Discussion

Principal Findings

We have obtained evidence that language models (particularly BERT-like models) can be used in preemptive mental health detection and analysis in longhand forums even if they have room for improvement.

In our preemptive detection depression experiment, depressed and controls were placed in ratios of 1:1, 1:3, 1:5, and 1:10. The purpose of which was to simulate increasingly realistic settings where most users are controls. In the balanced arrangement, 1:1, we obtained an F1 score and 0.738 using the MentalBERT LM. This is comparable to the works of Eichstaedt et al [14], De Choudhury et al [74], and Reece et al [19], which obtained F1 scores of 0.660, 0.680, and 0.650 respectively. This work provides evidence that LMs are more effective than existing methods for predicting depression in social media data in advance of diagnosis.

Our temporal analysis suggested that the final 12 weeks (about 3 months) of posts prior to a depressed user's estimated diagnosis date are likely to be most indicative of their condition. Another broader interpretation is that LMs do not appear to improve when adding more data prior to 12 to 16 weeks. BERT and MentalBERT obtained F1 scores of 0.726 and 0.715, respectively.

This contrasts to a certain extent with the results of Eichstaedt et al [14], albeit using AUC (Area Under Curve) scores rather than F1. Six months prior to the diagnosis date, 0.72 was obtained and three months prior 0.62 was obtained. From these results it is hard to draw clear conclusions, as the results may be affected by the nature of the data and models employed.

We also observed that posts made during the four-to-eight-week period prior to the user's estimated diagnosis date are also pertinent. They exhibit more negative sentiment than the posts made during any other four-week period (up to 24 weeks prior to their estimated diagnosis date). This finding may be supportive of prior work that distinct changes in mood may be predictive of the onset of depression [75].

We have been able to corroborate the significance of sentiment in the discourse of depressed users. We found that depressed users are approximately 1.18 times more likely to make a sentimental post than a non-depressed user.

Limitations

Constraints on our investigation primarily concern RSDD-Matched where its 70 depressed users make for a small sample. We did, however, use five-fold cross-validation to mitigate this and performed different experiments with various numbers of control users.

RSDD-Matched is derived from RSDD and RSDD-Time. As a result, the diagnosis dates of the users in RSDD-Matched are estimates only. Furthermore, posts made in mental health subreddits were deliberately elided from RSDD so were not available for consideration by our machine classifiers.

Conclusions

Using state-of-the-art language models, this work has posited how far in advance the diagnosis of depression in a person with depressive traits can be determined. With this knowledge, it may be possible to direct people with depression to a physician much sooner than they otherwise would. Moreover, perhaps more importantly, we have shown how these automatic NLP tools can serve to perform an analysis on the main traits arising from online postings.

We have also seen that the sentiment exhibited in the online forum postings demonstrates good sensitivity in detecting depressive traits.

Further work may include a multi-modal approach to the detection of people with depression in online forums, such as Reddit. For example, alongside the text of a Reddit users's posts, we might also consider the subreddits where they have upvoted and downvoted posts. Awards received or given may also be indicative of the user's mental health. Such a study would of course be contingent on the ability to synthesize a suitable dataset or source an existing one. Moreover, the usage of temporal information, such as Temporal Word Embeddings (TWE) [76], may enhance any multi-modal approach.

Methods for gauging the severity of depression in users of online forums may also be investigated. This might involve mining language features from user posts and observing how they correlate with ground-truth severity. Features of interest may include terms used in LIWC (Linguistic Inquiry and Word Count) dictionaries, sentiment, and emotion [77].

Data access statement

Information on the RSDD and RSDD-Time datasets used in this study, including their data access procedure, can be found at https://ir.cs.georgetown.edu/resources/rsdd.html.

Acknowledgements

Antonio F. Pardiñas was supported by an Academy of Medical Sciences "Springboard" award (SBF005\1083).
Jose Camacho-Collados is supported by a UKRI Future Leaders Fellowship.
We thank Professor Nazli Goharian of Georgetown University and Dr Andrew Yates of University of Amsterdam for their assistance in supplying RSDD and RSDD-Time.

Conflicts of Interest

None declared.

Abbreviations

ALBERT: A Lite BERT
BERT: Bidirectional Encoder Representations from Transformers
BioBERT: Bidirectional Encoder Representations from Transformers for Biomedical Text Mining

LDA: Latent Dirichlet Allocation
LM: language model
MDD: major depressive disorder
NLP: natural language processing
PMI: pointwise mutual information
RSDD: Reddit Self-reported Depression Diagnosis (dataset)
RSDD-Time: Reddit Self-reported Depression Diagnosis – time (dataset)
SVM: support-vector machine
TF-IDF: term frequency-inverse document frequency

References

1. Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx). URL: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b [accessed: 2021-05-01].
2. Kessler RC, Petukhova M, Sampson NA, Zaslavsky AM, Wittchen H -U. Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States. Int J Methods Psychiatr Res. 2012 Sep;21(3):169-84. doi: 10.1002/mpr.1359. Epub 2012 Aug 1. PMID: 22865617; PMCID: PMC4005415.
3. Regier DA, Kuhl EA, Kupfer DJ. The DSM-5: Classification and criteria changes. World psychiatry. 2013 Jun;12(2):92-8.
4. Picardi A, Lega I, Tarsitani L, Caredda M, Matteucci G, Zerella MP, SET-DEP Group. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. J Affect Disord 2016 Dec 01;198:96-101.
5. Edwards S, Tinning L, Brown JS, Boardman J, Weinman J. Reluctance to seek help and the perception of anxiety and depression in the United Kingdom: a pilot vignette study. J Nerv Ment Dis. 2007 Mar;195(3):258-61. doi: 10.1097/01.nmd.0000253781.49079.53. PMID: 17468687.
6. Wasserman C, Hoven CW, Wasserman D, Carli V, Sarchiapone M, Al-Halabí S, Apter A, Balazs J, Bobes J, Cosman D, Farkas L, Feldman D, Fischer G, Graber N, Haring C, Herta DC, Iosue M, Kahn JP, Keeley H, Klug K, McCarthy J, Tubiana-Potiez A, Varnik A, Varnik P, Ziberna J, Poštuvan V. Suicide prevention for youth--a mental health awareness program: lessons learned from the Saving and Empowering Young Lives in Europe (SEYLE) intervention study. BMC Public Health. 2012 Sep 12;12:776. doi: 10.1186/1471-2458-12-776. PMID: 22971152; PMCID: PMC3584983.
7. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: Proceedings of the 5th annual ACM web science conference 2013 May 2. p. 47-56.
8. Tadesse MM, Lin H, Xu B, Yang L. Detection of depression-related posts in reddit social media forum. IEEE Access. 2019 Apr 4;7:44883-93.

9. Malhotra A, Jindal R. Deep learning techniques for suicide and depression detection from online social media: A scoping review. Applied Soft Computing. 2022 Oct 14:109713.
10. Zhang T, Schoene AM, Ji S, Ananiadou S. Natural language processing applied to mental illness detection: a narrative review. NPJ digital medicine. 2022 Apr 8;5(1):1-3.
11. Yates A, Cohan A, Goharian N. Depression and self-harm risk assessment in online forums. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; 2017 Sep.
12. MacAvaney S, Desmet B, Cohan A, Soldaini L, Yates A, Zirikly A, Goharian N. RSDD-Time: Temporal annotation of self-reported mental health diagnoses. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic; 2018 Jun.
13. Owen D, Camacho-Collados J, Anke LE. Towards Preemptive detection of depression and anxiety in Twitter. In: Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task; 2020 Dec.
14. Abed-Esfahani P, Howard D, Maslej M, Patel S, Mann V, Goegan S, French L. Transfer Learning for Depression: Early Detection and Severity Prediction from Social Media Postings. InCLEF (Working Notes) 2019.
15. Shah FM, Ahmed F, Joy SK, Ahmed S, Sadek S, Shil R, Kabir MH. Early depression detection from social network using deep learning techniques. In2020 IEEE Region 10 Symposium (TENSYMP) 2020 Jun 5 (pp. 823-826). IEEE.
16. Eichstaedt JC, Smith RJ, Merchant RM, Ungar LH, Crutchley P, Preoţiuc-Pietro D, Asch DA, Schwartz HA. Facebook language predicts depression in medical records Proceedings of the National Academy of Sciences Oct 2018, 115 (44) 11203-11208; DOI: 10.1073/pnas.1802331115.
17. Ren L, Lin H, Xu B, Zhang S, Yang L, Sun S. Depression Detection on Reddit With an Emotion-Based Attention Network: Algorithm Development and Validation. JMIR Med Inform. 2021 Jul 16;9(7):e28754. doi: 10.2196/28754. PMID: 34269683; PMCID: PMC8325087.
18. Winokur G. Duration of illness prior to hospitalization (onset) in the affective disorders. Neuropsychobiology. 1976;2(2-3):87-93. doi: 10.1159/000117535. PMID: 1012452.
19. Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ. Forecasting the onset and course of mental illness with Twitter data. Sci Rep. 2017 Oct 11;7(1):13006. doi: 10.1038/s41598-017-12961-9. PMID: 29021528; PMCID: PMC5636873.
20. van Beljouw IM, Verhaak PF, Cuijpers P, van Marwijk HW, Penninx BW. The course of untreated anxiety and depression, and determinants of poor one-year outcome: a one-year cohort study. BMC Psychiatry. 2010 Oct 20;10:86. doi: 10.1186/1471-244X-10-86. PMID: 20961414; PMCID: PMC2974663.

21. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W. Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. J Affect Disord. 2014 Jan;152-154:45-51. doi: 10.1016/j.jad.2013.10.002. Epub 2013 Oct 21. PMID: 24183486.
22. Agorastos A, Marmar CR, Otte C. Immediate and early behavioral interventions for the prevention of acute and posttraumatic stress disorder. Curr Opin Psychiatry. 2011 Nov;24(6):526-32. doi: 10.1097/YCO.0b013e32834cdde2. PMID: 21941180.
23. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. 2017 Dec 1;18:43-9.
24. SMHD, RSDD, and RSDD-Time Datasets. Georgetown Information Retrieval Lab. URL: https://docs.google.com/forms/d/e/1FAIpQLScC-O3MXDd2lZSGqeRHsv1EMVR2xN5WC0cAodsHK3tBOz_FLw/viewform [accessed: 2020-11-21].
25. Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics. 2010 Dec;1(1):43-52.
26. Read J. Recognising affect in text using pointwise-mutual information. Unpubl. M Sc Diss. Univ. Sussex UK. 2004 Sep.
27. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information processing & management. 1988 Jan 1;24(5):513-23.
28. Lafon P. Sur la variabilité de la fréquence des formes dans un corpus. Mots. Les langages du politique. 1980;1(1):127-65.
29. Drouin P. Term extraction using non-technical corpora as a point of leverage. Terminology. 2003 Jan 1;9(1):99-115.
30. Camacho-Collados J, Pilehvar MT, Navigli R. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. Artificial Intelligence. 2016 Nov 1;240:36-64.
31. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory 1992 Jul 1 (pp. 144-152).
32. Pirina I, Çöltekin Ç. Identifying depression on reddit: The effect of training data. In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task 2018 Oct (pp. 9-12).
33. Salton G, Wong A, Yang CS. A vector space model for automatic indexing. Communications of the ACM. 1975 Nov 1;18(11):613-20.
34. Turian J, Ratinov L, Bengio Y. Word representations: a simple and general method for semi-supervised learning. In proceedings of the 48th annual meeting of the association for computational linguistics 2010 Jul (pp. 384-394).

35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011 Nov 1;12:2825-30.
36. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014 Oct (pp. 1532-1543).
37. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning 1998 Apr 21 (pp. 137-142). Springer, Berlin, Heidelberg.
38. Zhang W, Yoshida T, Tang X. Text classification based on multi-word with support vector machine. Knowledge-Based Systems. 2008 Dec 1;21(8):879-86.
39. Luss R, d'Aspremont A. Predicting abnormal returns from news using text classification. Quantitative Finance. 2015 Jun 3;15(6):999-1012.
40. Jardino M. Multilingual stochastic n-gram class language models. In: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings 1996 May 9 (Vol. 1, pp. 161-163). IEEE.
41. Vig J, Belinkov Y. Analyzing the structure of attention in a transformer language model. arXiv preprint arXiv:1906.04284. 2019 Jun 7.
42. Shen JH, Rudzicz F. Detecting anxiety through reddit. In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality 2017 Aug (pp. 58-65).
43. Burdisso SG, Errecalde ML, Montes y Gómez M. Using Text Classification to Estimate the Depression Level of Reddit Users. Journal of Computer Science & Technology. 2021 Apr 17;21.
44. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.
45. Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. 2019 Sep 26.
46. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234-1240. doi: 10.1093/bioinformatics/btz682. PMID: 31501885; PMCID: PMC7703786.
47. Beltagy I, Peters ME, Cohan A. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150. 2020 Apr 10.
48. Ji S, Zhang T, Ansari L, Fu J, Tiwari P, Cambria E. Mentalbert: Publicly available pretrained language models for mental healthcare. arXiv preprint arXiv:2110.15621. 2021 Oct 29.

49. Yu X, Hu W, Lu S, Sun X, Yuan Z. BioBERT based named entity recognition in electronic medical record. In 2019 10th international conference on information technology in medicine and education (ITME) 2019 Aug 23 (pp. 49-52). IEEE.
50. Alghanmi I, Espinosa-Anke L, Schockaert S. Interpreting patient descriptions using distantly supervised similar case retrieval. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval 2022 Jul 6 (pp. 460-470).
51. Bai Y, Zhou X. Automatic detecting for health-related twitter data with biobert. In Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task 2020 Dec (pp. 63-69).
52. González-Carvajal S, Garrido-Merchán EC. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012. 2020 May 26.
53. Clavié B, Alphonsus M. The Unreasonable Effectiveness of the Baseline: Discussing SVMs in Legal Text Classification. arXiv preprint arXiv:2109.07234. 2021 Sep 15.
54. Simple Transformers. URL: https://simpletransformers.ai/ [accessed: 2021-01-04].
55. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, Cistac P, Rault T, Louf R, Funtowicz M, Davison J. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771. 2019 Oct 9.
56. Classification Specifics – Simple Transformers. URL: https://simpletransformers.ai/docs/classification-specifics/#dealing-with-long-text [accessed: 2021-04-15].
57. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In Pacific-Asia Conference on Knowledge Discovery and Data Mining 2013 Apr 14 (pp. 201-213). Springer, Berlin, Heidelberg.
58. Hassan AU, Hussain J, Hussain M, Sadiq M, Lee S. Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. In 2017 international conference on information and communication technology convergence (ICTC) 2017 Oct 18 (pp. 138-140). IEEE.
59. Stephen JJ, Prabu P. Detecting the magnitude of depression in Twitter users using sentiment analysis. International Journal of Electrical and Computer Engineering. 2019 Aug 1;9(4):3247.
60. Liu T, Meyerhoff J, Eichstaedt JC, Karr CJ, Kaiser SM, Kording KP, Mohr DC, Ungar LH. The relationship between text message sentiment and self-reported depression. Journal of affective disorders. 2022 Apr 1;302:7-14.

61. Pota M, Ventura M, Catelli R, Esposito M. An effective BERT-based pipeline for Twitter sentiment analysis: a case study in Italian. Sensors. 2020 Dec 28;21(1):133.
62. Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.
63. Nguyen DQ, Vu T, Nguyen AT. BERTweet: A pre-trained language model for English Tweets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations 2020 Oct (pp. 9-14).
64. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823. 2014 Jun 23.
65. Wood SN, Pya N, Säfken B. Smoothing parameter and model selection for general smooth models. Journal of the American Statistical Association. 2016 Oct 1;111(516):1548-63.
66. The R Project for Statistical Computing. URL: https://www.r-project.org/ [accessed: 2022-07-13].
67. Matloff N. Statistical regression and classification: from linear models to machine learning. Chapman and Hall/CRC; 2017 Sep 19.
68. van der Brug W. Issue ownership and party choice. Electoral studies. 2004 Jun 1;23(2):209-33.
69. Guo H, Zhi W, Liu H, Xu M. Imbalanced Learning Based on Logistic Discrimination. Comput Intell Neurosci. 2016;2016:5423204. doi: 10.1155/2016/5423204. Epub 2016 Jan 4. PMID: 26880877; PMCID: PMC4736373.
70. Sokolova M, Japkowicz N, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. In Australasian joint conference on artificial intelligence 2006 Dec 4 (pp. 1015-1021). Springer, Berlin, Heidelberg.
71. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.
72. McCallum AK. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu. 2002.
73. Howes C, Purver M, McCabe R. Linguistic indicators of severity and progress in online text-based therapy for depression. In Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality 2014. Association for Computational Linguistics.
74. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. In Seventh international AAAI conference on weblogs and social media 2013 Jun 28.
75. van de Leemput IA, Wichers M, Cramer AO, Borsboom D, Tuerlinckx F, Kuppens P, van Nes EH, Viechtbauer W, Giltay EJ, Aggen SH, Derom C, Jacobs N, Kendler KS, van der Maas HL, Neale MC, Peeters F, Thiery E, Zachar P,

Scheffer M. Critical slowing down as early warning for the onset and termination of depression. Proc Natl Acad Sci U S A. 2014 Jan 7;111(1):87-92. doi: 10.1073/pnas.1312114110. Epub 2013 Dec 9. PMID: 24324144; PMCID: PMC3890822.

76. Couto M, Pérez A, Parapar J. Temporal Word Embeddings for Early Detection of Signs of Depression. In CIRCLE (Joint Conference of The Information Retrieval Communities in Europe), July 04-07 2022, Toulouse, France.
77. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology. 2010 Mar;29(1):24-54.