Forecasting mental health and emotions based on social media expressions during the COVID-19 pandemic

Anonymous for blind review requirement

Purpose. We present an approach for forecasting mental health conditions and emotions of a given population during the COVID-19 pandemic in Argentina based on social media contents.
Design. Mental health conditions and emotions are captured via markers, which link social media contents with lexicons. First, we build time series models that describe the evolution of markers, and their correlation with crisis events. Second, we use the time series for forecasting markers, and identifying high prevalence points for the estimated markers.
Findings. We evaluated different forecasting strategies that yielded different performance and capabilities. In the best scenario, high prevalence periods of emotions and mental health issues can be satisfactorily predicted with a neural network strategy, even at early stages of a crisis (e.g., a training period of 7 days).
Originality. Although there have been previous efforts to predict mental states of individuals, the analysis of mental health at the collective level has received scarce attention. We take a step forward by proposing a forecasting approach for analyzing the mental health of a given population at a larger scale.
Practical implications. This work contributes to a better understanding of how psychological processes related to crises manifest in social media, and this is a valuable asset for the design, implementation and monitoring of health prevention and communication policies.

Keywords: COVID-19; social media; psycho-linguistic text analysis; time series forecasting

1. Introduction

The COVID-19 pandemic has brought changes to people's social behavior, but also posed challenges for government and mental health. For instance, the World Health Organization has expressed concerns over the mental health and psycho-social consequences of both the pandemic and its preventive policies, which might increase loneliness, anxiety, depression, drug use, and suicidal behaviour among others (Kumar and Nayar, 2020). Over the last year, social media users have been massively expressing their thoughts and concerns regarding the pandemic. Furthermore, since social media are a bidirectional channel, the information posted can also affect these users and their mental health. This setting provides an opportunity for analyzing the pandemic effects on societal behaviors based on social media activity, and how such activity connects with existing theories about mental health and emotions. Moreover, as governments struggle to develop effective messaging strategies to support society, being able to analyze how society perceives and responds to those messages becomes crucial for decision makers. For example, the inclusion of altruism in UK official health messages has been reported to have a positive effect on well-being, in comparison with more imperative messages urging citizens to stay at home (Holmes et al., 2020).The analysis of textual exchanges in social media provides rich information about individuals' behaviors, which in turn sheds light on how a given crisis evolves, how individuals cope with the crisis, and how their mental health changes. The early identification of trends in mental health conditions of individuals (at the collective level) becomes crucial for decision makers when developing effective interventions or messaging strategies to support the affected population. For instance, during a period of high prevalence of tweets alluding to mental health (after 200 days of lock-down), the Argentinian authorities launched a campaign of mental health prevention both in social media and as part of diagnosing and contact tracing actions (Ministerio de Salud de Argentina, 2020).

In this work, we present an approach for forecasting mental health conditions and emotions of a given population during crises, like the COVID-19 pandemic, based on the language expressions in social media. Mental health conditions and emotions are captured via lexical categories, referred to as *markers*, which link social media contents to well-known lexicons. Our approach works in two stages. First, we construct descriptive charts for monitoring the *evolution of markers* and their correlation with crisis events or actions. As part of this stage, we detect *peaks* or *change-points* in the timelines, which represent time periods in which the observed markers undergo a substantial change with respect to previous observations. For example, high prevalence of anxiety and depression markers were observed around the time of the announcement of the first lock-down extension. Second, we model marker evolution as time series and support *marker forecasting* for a given time horizon. Markers can be combined into *dimensions* of analysis. This way, decision makers can assess what-if scenarios, and plan for possible interventions for the crisis. We studied mental health markers for three disorders (anxiety, depression and stress) and emotions (positive and negative) on a large collection of Twitter data related to the COVID-19 situation in Argentina. We explored both conventional time series and neural network strategies for forecasting high prevalence periods in the series, which would suggest a potential deterioration of mental health in the population.

We believe that this work contributes to a better understanding of the psychological processes related to crises as they reflect on Spanish-based social media and their users. Thus, this work has potential for informing the design of public policies for people's mental well-being during crises. Furthermore, the proposed approach is not tied to the Argentinian study, as it can be applied to other data streams or include other markers in the analysis.

The rest of the article is organized as follows. Section 2 presents background concepts about mental health as well as some related works. Section 3 describes our approach for deriving mental health markers based on existing sociological theories, and then forecasting the prevalence of mental health markers. Section 4 describes the study of the COVID-19 dataset using the approach. Finally, Section 5 presents the conclusions and discusses future lines of work.

2. Background and related work

Since the beginning of the COVID-19 pandemic (but also in previous crises), social media have become a rich information source for exposing the phenomenon, people's reactions and its effects. Besides disrupting life quality, crises often create a burden of mental health conditions by affecting individuals' expectations of the future, challenging their world view and even triggering emotional reactions (Kumar and Nayar, 2020). Hence, several quantitative analyses have studied mental health disorders (such as depression, suicidality and anxiety) and their symptomatology through natural language processing and psycho-linguistics techniques (Chancellor and Choudhury, 2020). For example, Gruebner et al. (2017) and Lin and Margolin (2014) aimed at identifying the basic emotions of Twitter users during Hurricane Sandy in 2012 and the Boston Marathon bombing in 2013, respectively. Gruebner et al. (2017) complemented the emotions analysis considering the geographic information of tweets to infer clusters of high emotion prevalence. Odlum and Yoon (2015) and Roy et al. (2020) focused on the Ebola epidemic in 2014. Odlum and Yoon (2015) explored emotion classification to determine whether public mood could be used for the early discovery of health threats. Emotion recognition was based both on lexicons and machine learning models. Results showed changes in the emotions expressed before and after an event, and according to the geographical closeness to the event. The authors concluded that social media could be used as a source of evidence for disease outbreak detection and monitoring. Roy et al. (2020), in turn, studied blame in social media. The authors performed a hand-coding thematic analysis of tweets and Facebook posts into different categories related to governments, media, affected populations and elite groups, among others. Results suggested an evolution of online blame from the affected populations to figures to which users showed pre-existing frustrations. In the data collected for this study, an evolution of blame was also observed. It started by blaming the people that returned from a trip abroad for the first contagions, then blaming the people enjoying sport activities for an increment of contagions, and lastly, manifestations of frustration against the government (referred to both health and economic situations).

As regards COVID-19, Li et al. (2020), Su et al. (2020) and Hou et al. (2020) explored Weibo to analyze the mental health impact of the pandemic based on the LIWC categories for emotions and concerns. Li et al. (2020) compared the prevalence of LIWC categories on a two-week period before and after the outbreak declaration. The study showed that negative emotions and sensitivity to social risks increased after the outbreak, while positive emotions decreased. Both Su et al. (2020) and Hou et al. (2020) focused on public emotion responses to epidemiological events, but also to government's announcements. According to Hou et al. (2020), anxiety, sad and anger peaks were reported after certain triggering events. Su et al. (2020) found significant differences in emotion prevalence in the different stages of the outbreak (initial, response, prevention and control). For example, while negative emotions had higher prevalence during the initial and response stages, higher prevalence was reported for positive emotions in the other two stages. Emotion recognition was based on lexicons. As shown by the works above, social media provide an opportunity for monitoring and analyzing stressing situations at a massive scale. Nonetheless, the analyses have been descriptive and retrospective in nature, as mental health states and trends were determined by looking at messages already shared by individuals. This limits the possibility of acting upon the crisis in the short term, for instance, by estimating possible health states that can alert about risky mental situations.

Several works have attempted to predict mental health disorders and well-being, based on individuals' activity information, including behavioral, mobility and sleeping patterns. For example, Umematsu et al. (2019) used deep learning models based on data extracted from wearable sensors, mobile phones, and behavioral surveys from a week to predict the stress level (self-reported by the individuals) of the next day. Also based on deep learning, Suhara et al. (2017) forecasted depression based on individuals' self-reports of activities and moods. Similarly, Reece et al. (2017) predicted depression diagnostics based on the presence of LIWC categories in social media posts. Predictions were based on a Hidden Markov Model and showed that three months prior to diagnosis, depressed subjects showed a marked rise in the probability of being in a depressed state, which decreased three months after the actual diagnosis. Both works focused on predicting depression at an individual level, instead of at a collective (or societal) level, as we proposed in this work.

Although there have been efforts in the literature to study the manifestation of mental health issues in social media, and some have tried to predict individuals' mental states, the analysis of mental health at the collective level has received less attention. In this context, we take a step forward in this direction by proposing a forecasting approach for collective trends based on a dataset of COVID-19 tweets.

3. Materials and methods

The proposed approach involves two stages: i) data preparation and operationalization of sociological theories in terms of markers, and ii) forecasting of markers and identification of their high prevalence periods. The first stage deals with data collection (e.g., tweets), pre-processing, and matching of the pre-processed data according to predetermined lexicons. This stage produces a series of charts showing the evolution of the *markers* and the mental health (i.e., `anxiety`, `depression`, `stress`) and emotions (i.e., `negative` and `positive emotions`) *dimensions* over time. A dimension is derived from a set of *markers* describing it, as schematized in Figure 1 (left side). Each *marker* represents a lexical category linking social media contents with the selected lexicons. On this basis, the second stage predicts the future values of markers (within a given time window) and then identifies peaks of high prevalence points for the corresponding dimension, as shown in Figure 1 (right side).

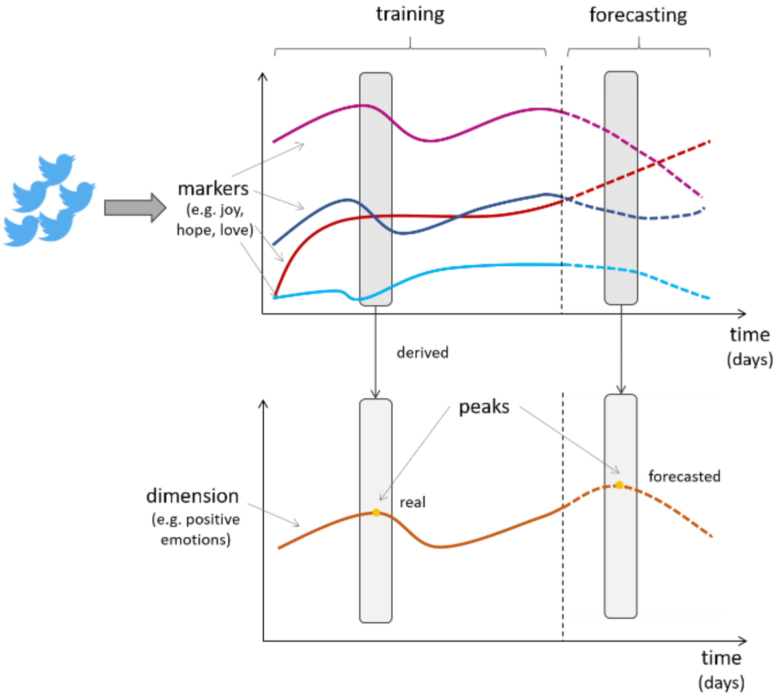


Figure 1. Evolution and forecasting of markers and dimensions

In the following sub-sections, we provide details of the approach and also the main research questions for our study.

3.1 Data collections and pre-processing

Our analyses are based on *SpanishTweetsCOVID19,* a large-scale sample of Twitter data collected during the COVID-19 pandemic in Argentina. *SpanishTweetsCOVID19* includes over 150 million tweets shared between March 1st and August 30th 2020 (Tommasel et al., 2020) [1]. Tweets were retrieved from the Twitter Streaming API in real time using the Faking it! [2] tool. This dataset provides a broad perspective of the mental health dynamics during the pandemic, with a center in the Argentinian situation. We collected tweets identified as written in Spanish, and including any of a set of keywords referring to COVID (e.g., `quedateencasa`, `covid`, `covid19`, `cuarentena`, `barbijo`, `salud`, etc.) or to official Argentinian offices or media (e.g., `msalnacion`, `alferdez`, `casarosada`, `infobae`, etc.), or included geographic information inside the Argentinian geographical bounding box. Retweets accumulated between the 15% (March) and 72% (May) monthly tweets. The number of retrieved tweets peaked in June. To keep tweets with enough content to evaluate mental health aspects, we only considered original tweets and replies. A data summary can be found in the companion repository [3]. Hashtags appeared only in the 19% of the kept tweets. A data inspection showed that neither politicians, media nor official government accounts included hashtags in most tweets. Collected tweets were pre-processed to remove URL, mentions and special characters. Hashtags were split into their constituent words. Spelling corrections were not applied as an inspection of randomly selected tweets showed that tweets were mostly correctly written. All pre-processing was implemented in Python using spaCy [4].

3.2 Lexicon-based analysis for markers

The language that users adopt in social media for expressing themselves provides evidence of their state of mind. Thereby, lexicons have been used for a text analysis of emotions, concerns, and health-related issues (De Choudhury et al., 2013). This study focuses on `positive` and `negative` emotions and three mental health dimensions: `anxiety`, `depression` and `stress`, which have been reported to increase during the COVID-19 outbreak (Kumar and Nayar, 2020). For the psycho-linguistic analysis, we rely on two lexicons: *SentiSense Affective* (de Albornoz et al., 2012) for emotions, and *Empath* (Fast et al., 2016) for mental health. The SentiSense Affective Lexicon labels WordNet synsets according to 14 emotional categories or *markers*. Markers comprise two dimensions: `positive` (calmness, joy, love, hope, like and anticipation) and `negative` (hate, anger, sadness, fear and disgust) emotions. As SentiSense is integrated with the WordNet Spanish version, it directly applies to the analysis of the collected tweets.

We analyzed each mental health dimension by matching the Empath categories with lexicons. Lexicons were derived from several literature sources: su Park (2012), Aldarwish and Ahmad (2017) and De Choudhury et al. (2013), as well as from the manual hand-coding of the characterizations of the disorders defined by the National Institute of Mental Health and the Anxiety and Depression Association of America (2020). These lexicons were semantically expanded using *FastText* [5] to capture the context in which words are used, and to understand how individuals express mental health related aspects (Losada and Gamallo, 2020). From these expansions, we automatically retrieved the top-10 Empath categories with the highest number of matching words with the lexicons, which represent the *markers* for the associated dimension. As Empath is only available in English, and given that the reliability of translating psycho-metric scales and lexicons has already been studied (Perczek et al., 2000), we translated it using IBM Watson [6]. These resources are provided in the companion repository [3].

The daily prevalence of a marker was computed as the percentage of tweets on such a day with at least a word matching the lexicon marker. Based on the marker prevalence in the whole time period, we model their sequence of observations as a *time series*. In our case, a time series models a sequence of observations about a given marker (or dimension) during a given time window. For emotions, the time period spanned between March-August, while for mental health it did not include July-August due to different magnitude orders across markers prevalence, which hindered the analysis. For identifying the high prevalence periods of a given dimension, we search the time series for *peaks*, which are points in which the markers for that *dimension* varied altogether, presumably due to COVID-related events. For example, the first `anxiety` peak matches the days following the confirmation of the first COVID case in Argentina and the first suspension of activities. For each *marker*, we computed the gradient over the one-week smoothed time series to reduce the impact of day-to-day variations, and favour consistent behaviors over longer time periods. Then, we averaged the gradients for all markers to obtain the overall gradient of the dimension, and computed the peaks. We only kept peaks with prominence values higher than the 80 percentile.

3.3 Forecasting markers and prevalence points

In time series, historical values can reveal the changing trend, which can be projected to future values of the series. Based on the history of mental health and emotion markers, we want to forecast their future values, which are later combined to derive the dimension values. From the dimension values, we can anticipate the emergence of high prevalence periods. For example, departing from the `anxiety` markers during the first two weeks of March, we can forecast the markers for the following week and compute the corresponding prevalence points for `anxiety`. To do so, three strategies were studied: i) univariate, ii) multivariate, and iii) deep learning.

Univariate forecasting

This is the simplest strategy, in which markers are separately forecast, disregarding possible dependencies between them. That is, markers are assumed to be independent from each other. We used ARIMA (Box and Jenkins, 1990) and Prophet (Taylor and Letham, 2018). ARIMA is a classical model that blends auto-regression (i.e., a predefined number of prior observations or lag order) and moving averages in stationary series. Thus, for our marker timelines, we previously checked whether the series were stationary, and if not, we de-trended them. Prophet is an additive regression model originally developed by Facebook that works best with series with seasonal effects and a large number of historical data. It is typically robust to shifts in the trend and handles outliers well. Unlike ARIMA, Prophet does not require transformations nor pre-processing to be applied to the data.

Multivariate forecasting

Mental health states might interact with each other, as supported by the notions of synchronization or response coherence (Kuppens and Verduyn, 2017). These interactions in the marker series cannot be captured with a univariate analysis. Multivariate analysis, in turn, models the dynamic relationships among a group of time series. A natural extension to univariate models is Vector Auto Regression (VAR) (Ltkepohl, 2007). VAR is a stochastic process that represents a group of time series as a linear function of their own past values (lag order) and the past values of all the other series in the group. This type of model has already been used in the forecast of emotions (Bringmann et al., 2018). As in ARIMA, series are required to be stationary and without seasonal trends. Also, the group of series should exhibit some correlation. A Granger causality test revealed indeed interactions among most markers for the analyzed dimensions.

Deep learning forecasting

Recurrent Neural Networks (RNN) (Williams and Zipser, 1989) are a multivariate strategy that is suited for learning problems with sequential data, as time series analysis, and can also capture complex dependencies. In RNN, the connections between the neurons can form cycles, building an internal memory that facilitates learning from naturally sequential data, in which new predictions depend on the previous ones. Predictions were based on a Gated Recurrent Unit (GRU) (Cho et al., 2014) network, a particular RNN that is effective for modeling varying length sequences and capturing short- to medium-range dependencies. GRUs have successfully been used for multivariate time series forecasting. They are less complex than other RNN, which translates into fewer parameters, and faster and better learning opportunities with limited data. The GRU was set to optimize the Mean Square Error loss with a RMSProp optimizer. Given that not every marker was on the same scale, data was normalized based on the quartile distribution. Like Prophet, no additional transformations were required.

3.4 Research questions

Based on the approach above, our goal was to assess whether time series forecasts were accurate enough, particularly for identifying high prevalence periods in the dimensions under analysis. We assumed that the temporal evolution of markers reflected people's emotions and mental health, and it was aligned with key pandemic-related events in Argentina. This assumption was empirically validated by the authors for the whole time period of the dataset. We addressed three research questions:

- **RQ1-** How does the quality of the forecast depend on the training period?
- **RQ2-** Does the forecast performance for mental health and emotional dimensions differ?
- **RQ3-** Which time series strategy does it provide the best forecast?

4. Data analysis and findings

For assessing the three strategies, we used a sliding window that consisted of a training period and a forecasting horizon, and rolled out the window over the whole timeframe of the series. The time period for observing mental health symptoms depends on the disorder under analysis. While anxiety and stress can manifest in the short-term, psychological questionnaires (such as PHQ-9 for depression) and guidelines (National Institute of Mental Health, 2020) state that symptoms can be present for at least two weeks before getting a diagnostic. Thus, for adequately capturing the mental health status, we varied the training period (for the time series strategies) between 7, 14 or 21 days. As time series forecasting often does not require long data sequences, we considered a time horizon of 7 days. As the experiments were not conducted in real time, we were able to evaluate the forecasting performance by comparing the predicted values against the real ones. Three data analyses were performed for evaluating the results with respect to the research questions. First, we looked at the forecasting errors of the strategies (ARIMA, Prophet, VAR and RNN) in terms of the Mean Absolute Percentage Error (MAPE) to determine their differences. The MAPE values for markers were compared using a paired statistical test, setting the p-value to 0.01, and quantifying the effect size. This comparison helped us to determine the optimal number of training days per strategy. For such optimal numbers, Table I summarizes the average forecasting errors for each dimension and their markers.

Table I. Summary of Mean Absolute Percentage Errors (MAPE) for the best performing forecastings

(a) Anxiety

	Anger	Confusion	Disappointment	Fear	Health	Horror	Nervousness	Sadness	Shame	Suffering
ARIMA	10.59 ±13.45	18.23 ±19.99	12.97 ±16.67	10.32 ±10.24	30.35 ±33.17	10.65 ±10.25	15.16 ±19.68	11.64 ±8.66	13.87 ±11.31	9.15 ±8.56
Prophet	11.91 ±18.79	15.89 ±23.15	15.01 ±16.16	13.8 ±16.74	34.1 ±42.47	18.89 ±32.88	19.27 ±32.63	13.66 ±9.52	13.83 ±12.27	9.69 ±11.33
VAR	8.88 ± 7.01	14.59 ± 16.13	13 ± 14.41	7.64 ± 5.91	15.23 ± 16.57	12.15 ± 13.66	7.79 ± 5.77	8.88 ± 7.11	14.99 ± 16.38	7.74 ± 6.23
Deep Learning	4.91 ± 4.9	6.69 ± 6.9	6.33 ± 7.85	7.2 ± 9.2	15.92 ± 14.45	8.41 ± 12.66	8.52 ± 10.69	5.91 ± 5.77	7.32 ± 7.48	5.83 ± 6.27

(b) Depression

	Disappointment	Disgust	Emotional	Neglect	Nervousness	Pain	Sadness	Shame	Suffering	Torment
ARIMA	12.97 ±16.67	11 ±12.8	19.25 ±46.52	11.97 ±12.15	15.16 ±19.68	9.26 ±11.55	11.64 ±8.66	13.87 ±11.31	9.15 ±8.56	16.69 ±17.94
Prophet	15.01 ±16.16	14.07 ±19.61	23.62 ±45.83	13.61 ±14.86	19.27 ±32.63	11.53 ±17.66	13.66 ±9.52	13.83 ±12.27	9.69 ±11.33	19.52 ±21.51
VAR	9.55 ± 9.8	11.94 ± 11.6	10.37 ± 9.39	11.05 ± 12.55	8.96 ± 8.54	9.73 ± 9.62	9.07 ± 11.79	10.32 ± 12.88	9.47 ± 11.08	10.55 ± 9.71
Deep Learning	8.17 ± 10.37	6.29 ± 7.83	6.54 ± 7.94	7.19 ± 7.74	7.36 ± 8.51	6.02 ± 8.37	4.85 ± 4.24	8.16 ± 9.62	4.87 ± 6.19	11.55 ± 17.02

(c) Stress

	Anger	Disgust	Fear	Health	Neglect	Nervousness	Sadness	Shame	Suffering	Torment
ARIMA	10.59 ±13.45	11 ±12.8	10.32 ±10.24	30.35 ±33.17	11.97 ±12.15	15.16 ±19.68	11.64 ±8.66	13.87 ±11.31	9.15 ±8.56	16.69 ±17.94
Prophet	11.91 ±18.79	14.07 ±19.61	13.8 ±16.74	34.1 ±42.47	13.61 ±14.86	19.27 ±32.63	13.66 ±9.52	13.83 ±12.27	9.69 ±11.33	19.52 ±21.51
VAR	25.14 ± 104.95	40.13 ± 81.51	37.59 ± 238.84	41.2 ± 86.01	59.19 ± 254.64	26.87 ± 143.85	312.24 ± 3146.58	24.61 ± 77.53	25.38 ± 75.8	19.48 ± 40.13
Deep Learning	4.68 ± 4.48	7.51 ± 8.33	8.03 ± 10.69	15.92 ± 13.38	7.73 ± 7.17	9.55 ± 11.62	5.48 ± 5.12	8.65 ± 8.6	5.03 ± 5.33	10.34 ± 16.85

(d) Positive Emotions

	Anticipation	Calmness	Hope	Joy	Like	Surprise
ARIMA	2.94 ± 3.82	7.09 ± 7.75	3.47 ± 4.37	4.76 ± 5.49	2.32 ± 3.34	4.26 ± 4.86
Prophet	3.48 ± 4.64	7.35 ± 7.39	4.46 ± 4.82	5.72 ± 6.48	3.43 ± 4.43	6.05 ± 6.03
VAR	2.98 ± 4.38	4.9 ± 5.8	3.63 ± 5.06	4.46 ± 5.88	2.96 ± 4.35	3.09 ± 4.45
Deep Learning	2.3 ± 3.44	6.85 ± 7.65	3.05 ± 3.84	5.04 ± 5.66	2.22 ± 2.75	4.07 ± 4.62

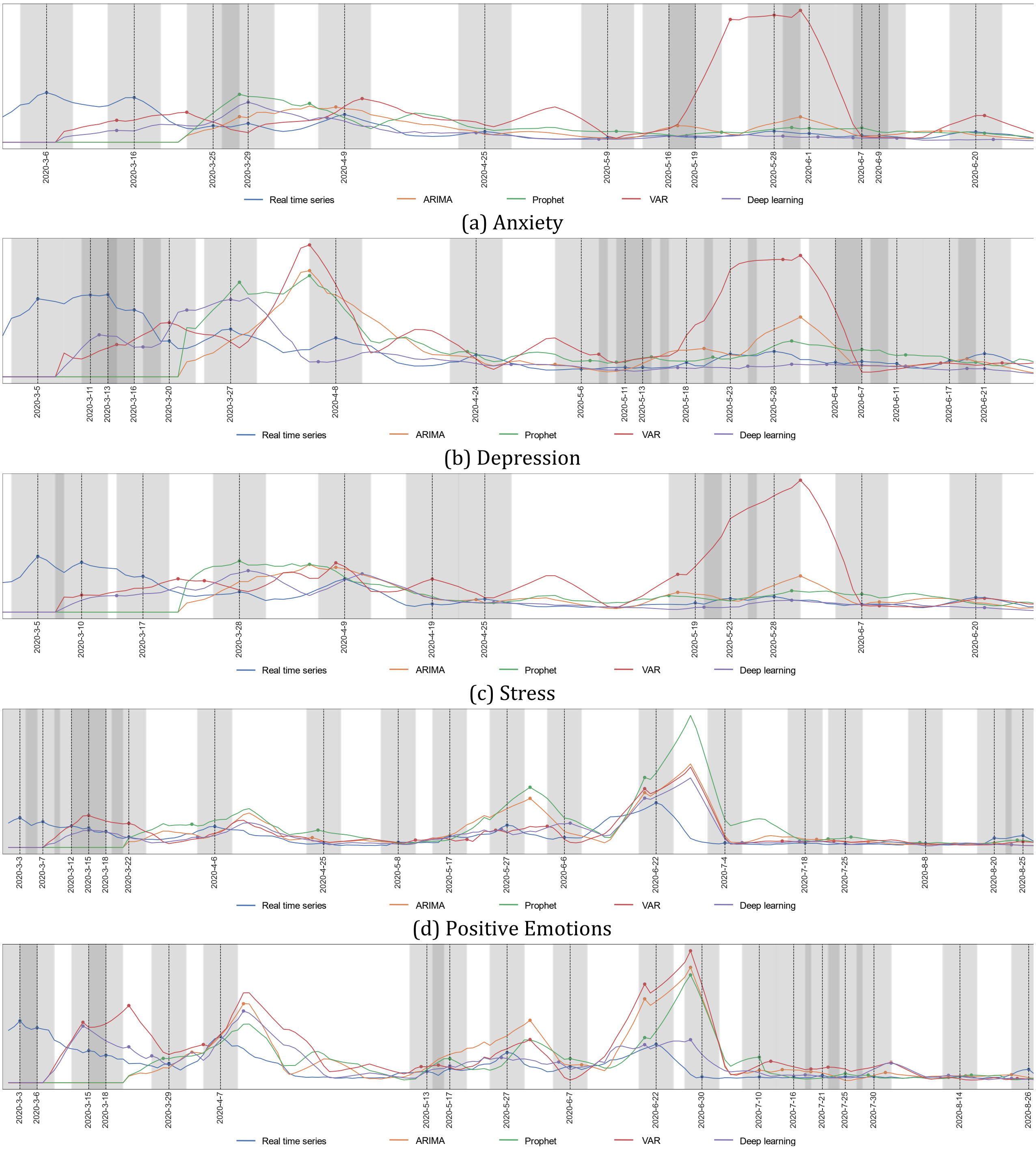
(e) Negative Emotions

	Anger	Disgust	Fear	Hate	Sadness
ARIMA	7.3 ± 8.07	2.54 ± 3.25	5.97 ± 6.85	14.03 ± 18.62	7.86 ± 7.2
Prophet	8.42 ± 7.98	3.17 ± 4.17	6.32 ± 6.55	13.24 ± 14.82	7.37 ± 6.78
VAR	6.93 ± 8.44	4.47 ± 5.54	3.95 ± 4.95	4.48 ± 5.85	8.51 ± 8.92
Deep Learning	8.01 ± 0	2.25 ± 2.84	6.3 ± 6.35	10.17 ± 16.25	7.48 ± 6.39

As a second step, we compared the real and forecast values for each marker using paired statistical tests. Ideally, the forecast and real values should not differ much from each other to allow a satisfactory prediction of high prevalence peaks. Lastly, we compared for each dimension the peaks resulting from the forecast markers and from the actual time series. To do so, we computed the hit-rate (or recall), which measures the proportion of true elements (i.e., peaks) discovered by the strategies. We considered a hit if a (predicted) peak was detected in a window of *n* days before and after an actual peak in the dimension time series, with *n* being 2, 3 or 7. For *n* = 7, a hit was successful if it was found in the same enclosing week. Table II presents the hit rates for the best performing forecasts (best results are in bold), and Figure 2 shows for a 3-day window (marked by the grey areas) the actual and forecast peaks for each dimension.

Table II. Hit Rates for the best performing forecastings for n= 2 - 3 – 7

	Anxiety	Depression	Stress	Positive Emotions	Negative Emotions
ARIMA	42% - 58% - 64%	21% - 50% - 57%	50% - 56% - 60%	44% - 52% - 63%	43% - 57% - 65%
Prophet	58% - 75% - 73%	40% - 60% - 67%	56% - 56% - 75%	33% - 44% - 59%	57% - 60% - 55%
VAR	50% - 67% - 73%	44% - 56% - 75%	55% - 64% - **78%**	47% - 63% - 57%	55% - **70%** - 57%
Deep Learning	**92% - 92% - 92%**	**83% - 89% - 92%**	**64% - 82% - 78%**	**50% - 67% - 71%**	**61% - 70% - 72%**



(a) Anxiety

(b) Depression

(c) Stress

(d) Positive Emotions

(e) Negative Emotions

Figure 2. Real and forecast prevalence and peaks for the analyzed dimensions

Next, we discuss the performance of each time series strategy and related findings. Extended versions of the tables can be found in the companion repository [3].

4.1 Univariate forecasting

Comparing the MAPE, we observed the lowest errors when considering 21 training days for both ARIMA and Prophet. In most cases, these errors were statistically significantly lower than those for 7 or 14 training days. For example, the `positive emotion` dimension showed differences for most markers, either using ARIMA or Prophet, while for the `negative emotion` dimension, differences were observed for all but the *hate* marker. Nonetheless, for certain markers and dimensions no significant error differences were observed. For example, using ARIMA the `anxiety` dimension had no difference for *sadness*, *nervousness* and *suffering*; while for Prophet differences were observed only for *disappointment*. We argue that the univariate analysis of markers is not enough for accurately forecasting them. The errors for `positive` and `negative emotions` were lower than those for the `mental health` dimensions, which implies that emotions are simpler to forecast, and that the inter-dependencies among such markers are weaker than those among mental health markers.

When comparing ARIMA and Prophet with 21 training days, we found small-to-medium significant differences favoring ARIMA. For example, significant differences were observed for every `positive emotion` but *calmness* and for three `negative emotions` (*hate*, *anger* and *fear*). Regarding the `mental health` dimensions, some differences favoring ARIMA were observed for `anxiety` (*sadness*, *fear*, *horror* and *disappointment*), `depression` (*sadness*, *emotional*, *disappointment*) and `stress` (*sadness* and *fear*). In general, Prophet performed worse than ARIMA, which could be because Prophet needs larger training periods.

The comparison between the gradient of the real and forecast markers revealed medium-to-large statistical differences for most markers. When comparing the gradients for the real and forecast dimensions, for ARIMA differences were small and even negligible. This could imply that although the marker forecasting is not perfect in terms of errors, it still captured the tendency of the dimension time series adequately (as shown in Figure 2). For Prophet, medium-to-large significant differences were observed in all cases. Regarding the hit rates in Table II, ARIMA and Prophet allowed to discover in average 40% and 49% of the peaks (n = 2), respectively. In average, the highest rates were observed for `negative emotions` (65% for ARIMA), and `anxiety` and `stress` (75% for Prophet). When considering the enclosing week for the hits (n = 7), the average hit rates increased to 62% and 66%, for ARIMA and Prophet, respectively. In summary, these results show that individually considering the markers is not enough for accurately forecasting the prevalence of mental health and emotions.

4.2 Multivariate forecasting

In VAR, the best results for each dimension did not follow the same trend as in the univariate case in which the best results shared the same training period (21 days). For `positive` and `negative emotions`, the lowest MAPE values were achieved with 7 training days, with similar results as those observed for ARIMA with 21 days. On the other hand, for `depression`, using 14 days led to a performance similar to that of ARIMA. In the case of `stress`, despite the average errors for the markers when using 7 days were higher than those for the other training days, small differences were observed only for two markers (*disgust* and *health*). This was caused by a highly skewed error distribution towards the high values, as shown by the standard deviations in Table I. Except for `stress`, in general, errors for VAR had a smaller range than those for ARIMA and Prophet. Regarding `emotions`, ARIMA and Prophet showed more stability in their results, which was evidenced by lower ranges in their forecasting. Minimum errors were also reduced by using VAR up to 95% (e.g., the *sadness* markers for `anxiety`).

The comparison between the real and forecast series revealed small-to-large differences for all markers, which is consistent with the univariate analysis results. Finally, as regards hit rate (Table II), VAR seemed to improve the results of ARIMA and Prophet for most dimensions. In average, considering a window of n=3 led to the discovery of 64% of the peaks. The highest improvements were observed for `depression` (104% regarding ARIMA) and `positive emotions` (42% regarding Prophet). As previously mentioned, despite a few days with high errors, the forecast series with 7 training days for `stress` maintained a shape similar to the original one, with an average hit rate improvement of 16% when compared to univariate analysis (n=7). For `positive emotions` hit rates were improved a 22% and 32% in average, for n=2 and n=3, while for `negative emotions` hit rates were improved by a 11% and a 20% in average, for n=2 and n=3, respectively.

The observed error differences for the same marker in the different dimensions (e.g., *disgust* belongs to both `depression` and `stress`) show the effect of the interrelations between markers in the forecast. Overall, multivariate results showed that considering the (implicit) relationships among markers improves the quality of forecasting, while reducing the training days needed.

4.3 Deep learning forecasting

The analysis of MAPE values for RNN showed the best results when using 7 training days. Moreover, errors were lower or similar to those observed when including more training days. The lack of significant differences implies that adding training days does not necessarily improve results. This observation agrees with Suhara et al. (2017), who stated that only the information of the last 14 days is relevant for forecasting psychological states.

When compared to the errors in ARIMA, Prophet and VAR, for the three `mental health` dimensions, we found medium-to-large statistically significant differences favoring the deep learning strategy. In this case, the largest error differences were in *emotional* (66%) for `depression`. Statistically significant differences favoring RNN were found for all markers. This means that the deep learning forecasting improved the best error levels of simpler techniques, while requiring less training data. A practical benefit of the RNN strategy is that enables decision makers to run forecasts and analyses sooner than with the other strategies. In the case of `emotions`, errors were already low for the best performing univariate and multivariate forecasts. Thus, the error variations with deep learning were smaller than those for `mental health`. In this regard, in some cases we did not observe significant differences with ARIMA. This trend was exhibited, for example, in *hate, like* and *love* for `negative` and `positive emotions`, respectively.

In several cases, we observed no significant differences between the real and the forecast marker values, which implies that the deep learning strategy can more adequately learn the marker trends. These results, in turn, led to better approximations of the real gradient of the dimensions (as shown in Figure 2) and higher hit rates. Table II shows that the deep learning forecast allowed to achieve the highest hit rates (with a $n = 3$), outperforming almost all univariate and multivariate hit rates. The two exceptions were `stress` and `negative emotions`, for which RNN achieved the same hit rate as VAR. The good performance of RNN can be attributed to the ability of neural networks to learn seasonal behaviors and complex (not linear) interactions among markers.

Finally, we answer our research questions as follows:

RQ#1 Results showed that the training periods achieving the best forecasts depended on the time series strategy. While simpler univariate forecasting required 21 days, for multivariate forecasting 7 days were often enough to improve the best univariate results. This could imply that simple strategies require more training days to compensate for disregarding the interactions between individual time series.

RQ#2 Differences were observed between the forecasts for the mental health and emotion dimensions. Emotions were simpler to forecast, yielding small errors even for the univariate strategies. This can be due to the interactions between emotion markers being not as relevant (for prediction) as those for mental health markers. Despite the multivariate improvements (over the univariate strategies) were lower than those for mental health markers, multivariate strategies still worked well with a 7-day training period.

RQ#3 The deep learning strategy achieved the highest quality forecasts, as they were able to leverage on the interactions among markers, and thus to capture the original trends (and peaks) of the dimensions. Moreover, using deep learning minimized the training period needed. Therefore, the emergence of high prevalence periods of emotions and mental health disorders could be anticipated even when limited data is available.

5. Conclusions

Crises, like the COVID-19 pandemic, affect society from multiple perspectives not only limited to physical health, but also to mental health, economics and politics. Hence, it is crucial to understand how people react to events and how their emotions and mental health evolve as crises develop. We believe that social media allows analyzing the mental health of a given population (or group of individuals) at a large scale.

From a practical point of view, the proposed approach provides a framework for forecasting mental health markers and estimating peaks with a high prevalence of (possible) mental health disorders in short-term time horizons. The framework is not tied to the specific markers and lexicons of the psycho-social theories used in our study, but rather it is general and it can be combined with other social media or psycho-social theories. We showed how different time series strategies offer different prediction capabilities for the task. In our experiments based on COVID-19-related tweets for Argentina, the forecasting based on neural networks provided the best results with a 7-day training period. In average, the neural network strategy detected 80% of the high prevalence peaks, with an average improvement of 48% regarding the univariate strategy. Differences were observed between the forecasts for the mental health and emotion dimensions, presumably due to the interactions among markers. In this sense, the deep learning forecasts better captured those marker interactions and thus the original trends of the time series.

Although the empirical work can benefit from a larger validation study, it shows evidence that the proposed approach can help in the design and monitoring of health prevention and communication policies (Holmes et al., 2020). In terms of psychological factors, the time series analyses could, in the short-term, support the monitoring of trends related to mental health issues, and the creation of different reports for health and government actors. The analyses can also foster the development of digital interventions (such as social media campaigns) to protect the mental well-being of citizens. To be genuinely effective, the intervention strategies need to be sensitive to the citizens' concerns, mental state and values (Hylanf-Wood et. al., 2021). Along this line, the current charts and analysis techniques should be enhanced to incorporate such information. Furthermore, causal relations between certain factors and the probability of peaks could be identified from the data, including contextual information for those relations. In the medium to long-term, the strategic definition of communication policies is fundamental for increasing the citizens' trust in them, in order to reduce the impact of the crisis. In turn, tracking the perceptions and responses of the designed campaigns would allow their iterative improvement and tailoring to the current societal situation.

From a social point of view, this kind of analyses can help understanding the role of social media consumption not only in potentially amplifying mental health issues, but also on affecting behaviour changes (e.g., the effect of social media campaigns explaining prevention measures, such as mask usage or hand sanitizing). The analyses could also foster the development of strategies for preventing the over-exposure to non-safe content, and encouraging citizens to stay informed by official channels in which they trust, which also helps mitigating the risk of misinformation and amplification of anxiety (Holmes et al., 2020).

We envision several aspects to be tackled as future works. First, this study focused so far on the society as a whole, without considering demographic differences among Twitter users. It would be interesting to segment users according to demographic characteristics (or other segmentation criteria) to observe how each group expresses during crises and how their mental health or emotions vary over time. This would enable the potential discovery of vulnerable groups (e.g., elderly, adolescents, people living alone, among others), so as to tailor the counteracting policies to their specific needs as required. For example, tailoring the vocabulary used in the communication of policies to a given target group, or even selecting different media for conveying the message. As an example of the latter, the Buenos Aires government designed a colourful graphic campaign and a song aiming at making it easier to remember the prevention measures during holidays (Ministerio de Salud de la Provincia de Buenos Aires, 2020) [7]. Second, we could compare how the pandemic manifested in neighbour countries and how the cultural dimensions and policies adopted by each government (in combination with its political orientation) might affect mental health manifestations. Third, in addition to Twitter, we could account for the perspectives provided by the users of other social media sites, and then assess how mental health manifests across them. At last, we intend to support decision makers by identifying which time events in the time series are the most influential for marker forecasting and peak discovery, based on recent developments on explainable Machine Learning for time series (Guillemé et al., 2019).

Notes

[1] Available at: https://data.mendeley.com/datasets/nv8k69y59d/2
[2] Available at: https://github.com/knife982000/FakingIt [3] Available at: http://bit.ly/3pFH8GS After review, this will be moved to a public repository.
[4] https://spacy.io/
[5] https://fasttext.cc/
[6] https://www.ibm.com/watson/services/language-translator/
[7] https://www.youtube.com/watch?v=8AuOBanae_w

References

Aldarwish MM and Ahmad HF (2017) Predicting depression levels using social media posts. In: 2017 IEEE 13th ISADS. pp. 277–280.

Anxiety and Depression Association of America (2020) Understanding the facts of anxiety disorders and depression is the first step. https://adaa.org/understanding-anxiety. (accessed October 20, 2020).Box GEP and Jenkins G (1990) Time Series Analysis, Forecasting and Control. USA: Holden-Day, Inc. ISBN 0816211043.

Bringmann LF, Ferrer E, Hamaker EL, Borsboom D and Tuerlinckx F (2018) Modeling nonstationary emotion dynamics in dyads using a time-varying vector autoregressive model. Multivariate behavioral research 53(3): 293–314.

Chancellor S and Choudhury MD (2020) Methods in predictive techniques for mental health status on social media: a critical review. Digital Medicine 3: 43.

Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H and Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: 2014 EMNLP. Doha, Qatar: ACL, pp. 1724–1734. DOI:10.3115/v1/D14-1179.

de Albornoz JC, Plaza L and Gervás P (2012) SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In: LREC'12. Istanbul, Turkey: ELRA, pp. 3562–3567.

De Choudhury M, Gamon M, Counts S and Horvitz E (2013) Predicting depression via social media. In: ICWSM. AAAI.

Fast E, Chen B and Bernstein MS (2016) Empath: Understanding topic signals in large-scale text. CHI '16. New York, USA: ACM, p. 4647–4657. DOI:10.1145/2858036.2858535.

Guillemé, M, Masson, V, Rozé, L and Termier, A (2019), November. Agnostic local explanation for time series classification. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 432-439). IEEE.

Gruebner O, Lowe SR, Sykora M, Shankardass K, Subramanian SV and Galea S (2017) A novel surveillance approach for disaster mental health. PLoS One 12(7).

Hyland-Wood, B, Gardner, J, Leask, J and Ecker, UK, (2021) Toward effective government communication strategies in the era of COVID-19. Humanities and Social Sciences Communications, 8(1), pp.1-11.

Holmes EA, O'Connor RC, Perry VH, Tracey I, Wessely S, Arseneault L, Ballard C, Christensen H, Silver RC, Everall I et al. (2020) Multidisciplinary research priorities for the covid-19 pandemic: a call for action for mental health science. The Lancet Psychiatry.Hou Z, Du F, Jiang H, Zhou X and Lin L (2020) Assessment of public attention, risk perception, emotional and behavioural responses to the COVID-19 outbreak: Social media surveillance in China. medRxiv DOI:10.1101/2020.03.14.20035956.

Kumar A and Nayar KR (2020) Covid 19 and its mental health consequences. Journal of Mental Health: 1–2.

Kuppens P and Verduyn P (2017) Emotion dynamics. Current Opinion in Psychology 17:22–26. DOI:10.1016/j.copsyc.2017.06.004. Emotion.

Li S, Wang Y, Xue J, Zhao N and Zhu T (2020) The impact of COVID-19 epidemic declaration on psychological consequences: A study on active Weibo users. Int. J. Environ. Res. Public Health. 17(6):2032.

Lin YR and Margolin D (2014) The ripple of fear, sympathy and solidarity during the Boston bombings. EPJ Data Science 3:31.

Losada DE and Gamallo P (2020) Evaluating and improving lexical resources for detecting signs of depression in text. Language Resources and Evaluation 54(1):1–24.

Ltkepohl H (2007) New Introduction to Multiple Time Series Analysis. Springer Publishing Company, Incorporated. ISBN 3540262393.Ministerio de Salud de Argentina (2020) El abordaje de la salud mental en el contexto de una pandemia sin precedentes https://www.argentina.gob.ar/noticias/el-abordaje-de-la-salud-mental-en-el-contexto-de-una-pandemia-sin-precedentes (accessed March 9, 2021).

Ministerio de Salud de la Provincia de Buenos Aires (2020) La provincia lanzó una campaña con un ABCD como fórmula de cuidados para evitar el aumento de casos de COVID-19 https://www.gba.gob.ar/saludprovincia/noticias/la_provincia_lanz%C3%B3_una_campa%C3%B1a_con_un_abcd_como_f%C3%B3rmula_de_cuidados_para(accessed March 9, 2021)

National Institute of Mental Health (2020) Mental disorders and related topics. https://www.nimh.nih.gov/health/topics/index.shtml. (accessed October 20, 2020).

Odlum, M, & Yoon, S (2015). What can we learn about the Ebola outbreak from tweets? American Journal of Infection Control, 43(6), 563–571. DOI:10.1016/j.ajic.2015.02.023

Perczek R, Carver CS, Price AA and Pozo-Kaderman C (2000) Coping, mood, and aspects of personality in spanish translation and evidence of convergence with english versions. Journal of personality Assessment 74(1):63–87.

Reece AG, Reagan AJ, Lix KL, Dodds PS, Danforth CM and Langer EJ (2017) Forecasting the onset and course of mental illness with twitter data. Scientific reports 7(1):1–11.

Roy M, Moreau N, Rousseau C, Mercier A, Wilson A and Atlani-Duault L. (2020) Ebola and Localized Blame on Social Media: Analysis of Twitter and Facebook Conversations During the 2014–2015 Ebola Epidemic. Culture, Medicine and Psychiatry 44: 56–79. DOI: 10.1007/s11013-019-09635-8

Su, Y, Wu, P, Li, S, Xue, J, and Zhu, T (2020). Public emotion responses during COVID-19 in China on social media: An observational study. Human Behavior and Emerging Technologies. DOI: 10.1002/hbe2.239

su Park M (2012) Exploring healthcare opportunities in online social networks: Depressive moods of users captured in twitter. In: ACM SIGKDD HI-KDD.

Suhara Y, Xu Y and Pentland AS (2017) Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In: WWW '17. p. 715–724. DOI:10.1145/3038912.3052676.

Taylor SJ and Letham B (2018) Forecasting at scale. The American Statistician 72(1): 37–45.

Tommasel A, Rodriguez JM and Godoy D (2020) SpanishTweetsCovid-19: A social media enriched COVID-19 Twitter Spanish dataset. DOI:10.17632/nv8k69y59d.1.