



# Public wellbeing analytics framework using social media chatter data

Heba Ismail<sup>1</sup> · M. Adel Serhani<sup>2</sup> · Nada Hussien<sup>1</sup> · Rawan Elabyad<sup>1</sup> · Alramzana Navaz<sup>2</sup>

Received: 7 August 2022 / Revised: 28 September 2022 / Accepted: 13 October 2022 / Published online: 3 November 2022  
© The Author(s) 2022

## Abstract

Public wellbeing has always been crucial. Many governments around the globe prioritize the impact of their decisions on public wellbeing. In this paper, we propose an end-to-end public wellbeing analytics framework designed to predict the public's wellbeing status and infer insights through the continuous analysis of social media content over several temporal events and across several locations. The proposed framework implements a novel distant supervision approach designed specifically to generate wellbeing-labeled datasets. In addition, it implements a wellbeing prediction model trained on contextualized sentence embeddings using BERT. Wellbeing predictions are visualized using several spatiotemporal analytics that can support decision-makers in gauging the impact of several government decisions and temporal events on the public, aiding in improving the decision-making process. Empirical experiments evaluate the effectiveness of the proposed distant supervision approach, the prediction model, and the utility of the produced analytics in gauging the public wellbeing status in a specific context.

**Keywords** Public wellbeing · Social media · Twitter · Emotion lexicon · Distant supervision · Sentence embeddings · Spatiotemporal analytics

## 1 Introduction

The increase in complexity and pace of modern life worldwide has contributed to a higher percentage of wellbeing issues. Governments spend a fortune on supporting public wellbeing. According to the Global Burden of Disease (GBD) Study 2017, the global economic burden of mental diseases was estimated to be \$ 99.5 million (Khaledi et al. 2019). Public wellbeing is undeniably affected by several

life events. An evident occasion representing this significant impact is the outbreak of COVID-19 in March 2020. The unprecedented global outbreak of COVID-19 has caused governments worldwide to force several strict disease control measures (Li et al. 2020a, b). These control measures resulted in increased negative emotions in public, hindering the public's wellbeing (Kydros et al. 2021; N et al. 2021). Even after the pandemic's peak, these negative consequences on public wellbeing still hold so far. In an attempt to assess public wellbeing, many governments have conducted several country-wide surveys to assess the public's wellbeing (Allen et al. 2021; Shahriarirad et al. 2021; Son et al. 2020; Thapa et al. 2020). However, the public wellbeing surveys used are deficient in inferring a holistic insight into the public's wellbeing due to the content limitation (i.e., intensive use of closed-ended questions), the subjectivity of the analysis, time limitations, geographic limitations, and other factors (Jones et al. 2008; Schwartz et al. 2014). This has necessitated a more holistic framework, adaptive to variable contexts and semantics, to assess the public wellbeing problems associated with various events. Social media platforms can give a representative and semantically-rich insight into the public's wellbeing. During the pandemic and on different other occasions such as elections, new year,

✉ Heba Ismail  
heba.ismail@adu.ac.ae

M. Adel Serhani  
serhanim@uae.ac.ae

Nada Hussien  
1071042@students.adu.ac.ae

Rawan Elabyad  
rawan.elabyad@adu.ac.ae

Alramzana Navaz  
201570182@uae.ac.ae

<sup>1</sup> College of Engineering, Abu Dhabi University, Abu Dhabi, UAE

<sup>2</sup> College of IT, United Arab Emirates University, Al Ain, UAE

holiday seasons, or war, people find social media as a free platform to exchange opinions, express emotions and discuss related concerns (Dubey 2020). Twitter among other social media platforms is seen as the most popular and accessible social media platform for sharing public concerns (Kydros et al. 2021). In several studies, Twitter was considered a reliable source revealing wellbeing indicators (Boon-Itt & Skunkan 2020; De Choudhury et al. 2013). Therefore, this study targets Twitter owing to its popularity and accessibility; however, the proposed framework applies to other social media platforms.

To this end, we propose an end-to-end public wellbeing assessment framework that integrates semantically-rich sentence embeddings with predictive models to provide spatiotemporal, real-time analytics on public wellbeing using unstructured data found on social networking sites. Wellbeing indicators such as emotional orientation, the level of public interaction with positive and negative opinions, and emotional intensity relevant to specific events, in specific locations, and over variable timeframes will be analyzed in the proposed framework. Sentence embeddings are generated using the Pretrained Deep Bidirectional Transformer Network (BERT) (Devlin et al. 2019). Wellbeing training datasets in the proposed model are produced regularly using a novel distant supervision approach to provide context-relevant training data.

The proposed framework overcomes many challenges reported in the literature on automatic wellbeing assessment from the unstructured text found on social media. For instance, the proposed framework implements a novel heuristic-based, distant supervision method using an expert's-verified wellbeing lexicon, providing a current and semantically relevant wellbeing training dataset. Wellbeing indicators in unstructured text vary depending on the linguistic context. User-generated unstructured text on social media differs over time, covering various topics of interest influenced by several life events. As such, historical training data used in previous research studies might not bear semantic relationships relevant to the current context, making them less valuable in training predictive models for a different timeframe (Almouzini et al. 2019; Chatterjee et al. 2019; Coppersmith et al. 2015; Scheuer et al. 2020).

In addition, manual annotation approaches commonly used in the literature on wellbeing assessment are time-consuming and inclined to subjectivity (Almouzini et al. 2019; Chatterjee et al. 2019; Coppersmith et al. 2015; Scheuer et al. 2020). Therefore, in this study, a wellbeing-specific distant supervision method is proposed to produce current training datasets bearing semantic relationships to emerging contexts. This alleviates the need for time-consuming manual annotation. Moreover, the proposed framework overcomes the feature engineering challenge inherent in the unstructured text stated by several studies in the field

of wellbeing assessment (Almouzini et al. 2019; Chatterjee et al. 2019; Coppersmith et al. 2015; Scheuer et al. 2020) by regularly generating semantically-rich sentence embeddings using BERT from current social media content. In that sense, the proposed framework is adaptive to the changes in the semantic relationships and contextual features in the user-generated unstructured content, providing a more accurate assessment of wellbeing status. For further validation, we present a use case from the education context to demonstrate the utility of the proposed framework in inferring wellbeing insights related to COVID-19.

This paper is organized as follows: first, we review and summarize the literature on automatic wellbeing assessment from unstructured text, highlighting the several challenges addressed by other research studies, and then we describe the proposed framework and its main components. Afterward, we conduct a series of experiments to validate the proposed distant supervision method, the wellbeing predictions, and finally, the utility of the proposed framework in a particular context.

## 2 Literature review

This section reviews related literature in the field of automated wellbeing assessment using social media content. We start by reviewing literature related to public wellbeing assessment highlighting limitations of existing research, and then we introduce the challenges related to feature engineering, contextual language models, and the inadequacy of training datasets for wellbeing assessment. The reviewed research studies are summarized in Table 1.

### 2.1 Analyzing the public wellbeing

Understanding the emotional status of the public across various instantaneous and ever-changing events is of utmost priority to all countries. Therefore, several studies investigated public wellbeing and mental health status through the automatic analysis of social media platforms or through surveying techniques. The reported studies on automatic assessment of public wellbeing using social media content focused mainly on emotions and sentiment analysis of user-generated text in specific locations. The most recent studies analyzed the influence of COVID-19 on public wellbeing in Iran, India, Norway, Sweden, the USA, and Canada (Bustos et al. 2022; Li et al. 2020a, b; Shahriarirad et al. 2021), China (Li et al. 2020a, b; Zhou et al. 2020), as well as Greek (Kydros et al. 2021). While the majority focused on predicting positive and negative sentiments, some studies analyzed distinct emotions such as Fear, Sadness, and Disgust. The reported studies do not conduct thorough spatiotemporal analysis and employ manually annotated datasets from previous contexts.

**Table 1** Summary of relevant research works focused on analyzing wellbeing from Twitter content highlighting features, annotation method, and prediction accuracies

| Work                         | Language | Method                                     | Features         |                    |                  |        | Ground Truth Annotation |             |          |                 |
|------------------------------|----------|--|------------------|--------------------|------------------|--------|-------------------------|-------------|----------|-----------------|
|                              |          |  | Machine Learning | Deep Learning      | BoW              | Topics | Emotion Terms           | Behavioural | Metadata | Embeddings      |
| Tsugawa et al. (2013)        | Japanese | Regression (Reg Coef 0.43)                 | ✓                |                    |                  |        |                         |             |          | SDS             |
| Resnik et al. (2015)         | English  | Support Vector Regression                  | ✓                | LDA                | LIWC             |        |                         |             |          | ✓               |
|                              |          | Precision (62%–74%)                        |                  | SLDA<br>SNLDA      |                  |        |                         |             |          |                 |
| Resnik et al. (2013)         | English  | Linear Regression                          |                  | LDA                | LIWC             |        |                         |             |          | Big-5 BDI       |
|                              |          | Precision (43%–50%)                        |                  |                    |                  |        |                         |             |          |                 |
| Tsugawa et al. (2015)        | Japanese | Support Vector Machines Accuracy (61%–66%) | ✓                | LDA                | Japanese Lexicon | ✓      | ✓                       |             |          | CES-D BDI       |
| De Choudhury et al. (2013)   | English  | Support Vector Machines Accuracy (70%)     |                  | LIWC               | ✓                | ✓      |                         |             |          | CES-D BDI       |
| Reece et al. (2016)          | English  | Random Forest–ROC (0.87–0.89)              |                  | LIWC A-NEW         | ✓                | ✓      |                         |             |          | CES-D TSQ       |
|                              |          |  |                  |                    |                  |        |                         |             |          | ✓               |
| Almouzini et al. (2019)      | Arabic   | Multiple Classifiers Accuracy (55%–87%)    | ✓                |                    |                  |        |                         |             |          | CES-D PHQ-9     |
| Shetty et al. (2020)         | English  | Multiple Classifiers Accuracy (72%–76%)    |                  | LSTM & CNN         | ✓                |        |                         |             |          |                 |
|                              |          |  |                  | Accuracy (93%–95%) |                  |        |                         |             |          |                 |
| Husseini Orabi et al. (2018) | English  | SVM –Accuracy (73%–77%)                    |                  | CNN & RNN          | ✓                |        |                         |             |          | Word2Vec        |
|                              |          |  |                  | Accuracy (51%–87%) |                  |        |                         |             |          | CBOW Skip-grams |
|                              |          |  |                  |                    |                  |        |                         |             |          | ✓               |

On the other side, several studies analyzed public wellbeing through surveying techniques. For instance, Patrick et al. (2020) conducted a national survey from June 5 to 10, 2020, to assess the COVID-19 pandemic and mitigation efforts that affected the physical and emotional wellbeing of parents and children in the USA. However, the collected data was at a one-time point rather than longitudinally. Also, the survey results were biased since the respondents had higher socio-economic status levels than non-respondents. Another study by VanderWeele et al. (2021) investigated the impact of the COVID-19 pandemic on physical and mental health from January to June 2020. The study employed a 15-min questionnaire, and the participants were selected using a stratified national sample of adults 18 and older within the USA. The study investigation was constrained by the sampling technique, the online assessment's requirement for internet access, the potential seasonality of respondents' wellbeing, and the potential for selection bias given that respondents in June were less likely to participate than in January. Krendl & Perry (2021) examined older adults' mental health and social wellbeing in a short time to measure social isolation due to COVID-19. Between April and May 2020, around 93 older adults in the USA took place on phone interviews conducted by the authors. Moreover, a study by Wang et al. (2020) analyzed the mental health outcomes associated with the covid-19 pandemic. They used online surveys of the Chinese population with a snowball sampling design. The survey took place between January and February 2, 2020, with a total of 1210 respondents from 194 cities in China. Notwithstanding, the selection bias resulted from an oversampling of a specific network of peers. Therefore, the finding was less applicable to the overall population, especially those with lower levels of education. Therefore, to our best knowledge, the existing studies in the field of public wellbeing analysis are usually designed for a specific timeframe, location, or sample, or rely on manually annotated data from different contexts. None of the reported studies provided an end-to-end framework that starts with generating a wellbeing-specific labeled dataset leading to inferring spatiotemporal insights across variable timeframes and geo-locations.

## 2.2 Feature engineering challenges in machine learning-based methods for wellbeing assessment

Several research efforts in automated wellbeing assessment on social media focused on building classical machine learning predictive models, which relied on "features" (Resnik et al. 2013, 2015; Tausczik & Pennebaker 2010). These studies focused mainly on linguistic features extracted from user-generated text. For instance, Tsugawa et al. (2013) utilized frequency bag-of-word (BoW), while Resnik et al. (2015), and Armstrong (Resnik et al. 2013), used various

combinations of topic models. In addition, some studies tried to improve on results obtained from linguistic features by extracting terms with emotional sentiment based on emotions' lexicons, such as the Linguistic Inquiry Word Count (LIWC) Lexicon (Tausczik & Pennebaker 2010). For example, Schwartz et al. (2014) utilized some emotion lexicons, including LIWC, to generate several feature vectors. A major challenge inherent in using linguistic features of all types is the very high dimensionality of the feature space, which usually leads to sparse datasets. Therefore, various features selection techniques have been applied to obtain the most prevalent features, such as correlation analysis (Tsugawa et al. 2013), attempting different classification thresholds (Resnik et al. 2015), or even involving a clinical psychologist to manually evaluate the extracted features (Resnik et al. 2013), a task which is expensive and time-consuming.

Besides the feature selection challenge, reported prediction accuracy, relying barely on linguistic features, was not very promising, as depicted in Table 1. Consequently, several studies attempted different behavioral and metadata features, which were argued in the literature to be more prevalent in wellbeing issues. For instance, Tsugawa et al. (2015) used, in addition to frequency bag-of-word (BoW) and Latent Dirichlet Allocation (LDA) topics, the ratio of positive-affect and negative-affect words based on a Japanese affect dictionary (Inui et al. 2013) Furthermore, they analyzed behavioral features such as tweet time, hashtags, and URLs indicating some symptoms of depression such as insomnia or isolation. Choudhury et al. (2013) explored even more sophisticated features by looking at the use of depression-related vocabulary based on terms extracted from Yahoo Answers under the "Wellbeing" Category. They also counted the frequency of names of antidepressants using a "list of antidepressants" from the Wikipedia page and extracted egocentric features based on social network analysis; (a) node properties, (b) dyadic properties, and (c) network properties. Reece et al. (2016) explored the combination of multiple emotion lexicons with a more comprehensive set of behavioral and metadata features in addition to time series features. Even though prediction results improved slightly with the addition of behavioral and metadata features, complexity increased too, adding more overhead and making these methods very expensive and time-consuming.

Moreover, to accurately extract and validate all these features, one must have adequate knowledge relevant to wellbeing research making these approaches unsuitable for continuous public wellbeing assessments. Moreover, latent semantics and contextual linguistic signs are key indicators of wellbeing and other emotional states which cannot be ignored. Nevertheless, features extracted from text, or other types of behavioral features, lack context. Consequently, it has become imperative to employ efficient prediction models

that do not require explicit feature extraction or identification and modeling of contextual attributes.

### **2.3 Contextual language models' limitations in deep learning-based wellbeing assessment methods**

Deep Learning (DL) algorithms integrating automatic feature representation have made impressive advances in the field of natural language processing (NLP), alleviating the inherent limitations in classical machine learning algorithms that depend on hand-crafted features (Young et al. 2018). For instance, Shetty et al. (2020) used raw twitter data with two Deep Learning algorithms; long short-term memory (LSTM) and sequential Convolutional Neural Network (CNN). They compared results with multiple machine learning classifiers trained on TF-IDF and frequency BoWs. Results of Deep Learning classifiers surpassed those of machine learning classifiers reducing the overhead associated with feature extraction and selection. Moreover, Orabi et al. (2018) investigated the use of multiple word-embeddings. They used three word embeddings; Word2Vec, CBOW, and Skip-grams, with CNN and Recurrent Neural Network (RNN), compared to SVM trained on TF-IDF BoW. Results obtained from DL predictive models suggest that DL models trained on word embeddings can outperform the prediction accuracy of classical machine learning and DL models trained on discrete word vectors for wellbeing assessment. Word vectors overcome some challenges inherent in features engineering highlighted earlier. However, they suffer from three main limitations. Word vectors are context-independent, i.e., (1) do not reflect contextual relationships among words, (2) do not account for unknown words, i.e., they do not support transfer-learning to a model using a different vocabulary of the same size, and (3) the size of vectors depends on the size of vocabulary used, which can be huge, resulting in the curse of dimensionality problem. Nevertheless, context can reveal useful wellbeing indicators. Hence, in this research, we intend to address the feature engineering challenges explained above and the context-aware challenge by implementing the Pre-trained Deep Bidirectional Transformer Network (BERT) to regularly generate fully contextualized sentence embeddings relevant to different life events reflecting a more relevant representation of wellbeing linguistic markers.

### **2.4 Inadequacy of existing labelled training data for wellbeing**

Likewise, the feature engineering and context modeling challenges highlighted earlier, acquiring accurate and representative labeled training data is a persisting challenge. Wellbeing datasets obtained from social media platforms

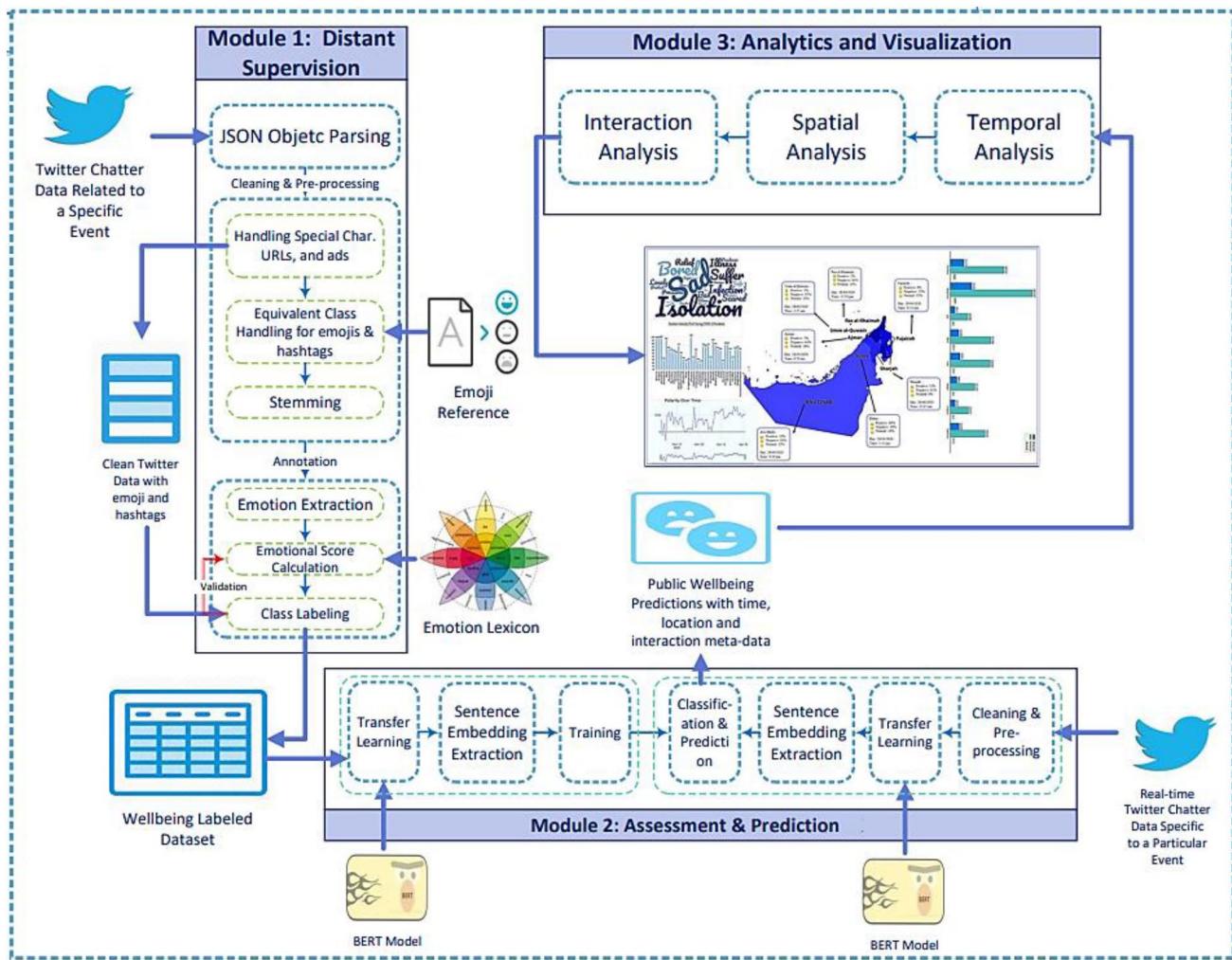
already exist. However, they are specific to previous timeframes and reflect different contexts and topics of interest. For instance, there exist twitter datasets focusing on public wellbeing post-COVID-19 (Kleinberg et al. 2020). Moreover, some datasets were collected after explicitly asking participants to express and reflect on a particular topic, for example, (Scheuer et al. 2020). Furthermore, some are solely focused on emotion detection only (Chatterjee et al. 2019) or diagnosing certain mental health disorders such as depression (Almouzini et al. 2019) and post-traumatic-stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD) (Coppersmith et al. 2015).

Nevertheless, as illustrated in Table 1, annotation approaches in the field of wellbeing assessment classically relied on manual annotation by (i) wellbeing experts (Resnik et al. 2015), (ii) psychometric questionnaires (e.g., CES-D & BDI) through crowdsourcing (Tsugawa et al. 2015), (De Choudhury et al. 2013), or (iii) self-reporting (Guntuku et al. 2017). These labeling methods cannot be used under different circumstances in various contexts. We summarize the inherent limitations hereafter:

- 1 They are expensive, time-consuming, and rely solely on manual annotation. With the ever-evolving number of Tweets related to wellbeing, we need a reliable automated labeling approach capable of annotating current, real-time data with minimal manual intervention to reflect the specific semantics relevant to this data.
- 2 Available wellbeing training datasets focus on patients diagnosed with a mental disorder. In previous research, identified users' tweets are usually tracked for one year (De Choudhury et al. 2013; Tsugawa et al. 2015). We are looking for early indicators to understand public wellbeing status and support people early on before developing mental disorders.
- 3 Existing labeled data for wellbeing are outdated and collected from different contexts. We need timely labeled data to analyze emotional triggers relevant to the public's wellbeing for different spatiotemporal events and provide more relevant support.

## **3 Methods**

To address the above-mentioned research challenges and achieve the research aims, we propose a framework composed of three main modules as illustrated in Fig. 1: (1) Distant Supervision, (2) Assessment and Prediction, and (3) Analytics and Visualization. The three modules collaborate to infer real-time insights into public wellbeing indicators to support decision-makers plan, design, and



**Fig. 1** Public wellbeing analytics framework

implement appropriate interventions. In subsequent sections, we briefly describe the tasks carried out in each module.

### 3.1 Module 1: distant supervision

This module is designed to generate current wellbeing-related labeled datasets based on distant supervision. The context changes with the continuously changing situations. Survey-based (De Choudhury et al. 2013; Tsugawa et al. 2015) and expert-based (Resnik et al. 2015) approaches, commonly used in the field of automatic wellbeing assessment, are not practical nor ideal solutions for rapidly changing contexts. Also, historical ground truth benchmark data generated from other contexts might not bear semantics and topics related to new and emerging situations, influencing individuals' emotional states. This, in turn, hinders the accuracy of Predictions (Li et al. 2020a, b). Hence, we define and validate a fully automated, distant supervision approach

suitable to generate up-to-date labeled data for any given situation or context based on verified emotion lexicons and expert-verified heuristics. This module's first, anonymized and representative, raw Twitter chatter data is hydrated in JSON format. Then, several subprocesses are carried out to ensure accurate annotation of data. These are: (1) parsing JSON objects' attributes required for wellbeing prediction as well as for analytics modules, (2) cleaning and pre-processing tweet's text to prepare it for annotation, and then (3) annotating the cleaned and pre-processed tweets based on emotional lexicons such that a total emotional score is calculated based on a predefined function. Hereafter, we briefly describe each one of these three sub-processes.

#### 3.1.1 Parsing JSON objects

All Twitter APIs used to retrieve Twitter data encode this data using JavaScript Object Notation (JSON). JSON is based on key-value pairs with named attributes and

associated values. These attributes and their state are used to describe objects. Twitter defines the number of JSON objects for each account. We focus primarily on the “Tweet Object” and the “User Object” to extract attributes such as ‘text,’ ‘place,’ ‘hashtag,’ ‘favorite\_count,’ and ‘retweet\_count’. The “Text” attribute is primarily used to retrieve the user’s generated text and assess the use of negative language in addition to other insights. Other attributes are essential for the Analytics Module.

### 3.1.2 Cleaning and pre-processing tweet’s text

Twitter is full of peculiarities; on the other hand, these special characters, emojis, and hashtags might reveal valuable semantics relevant to users’ wellbeing, as indicated by several studies, e.g. (Alvarez-Jimenez et al. 2019). Click or tap here to enter text. Hence, we conduct pre-processing tasks such as tokenization, stemming, and sentiment replacement in two phases. Initially, we generate a clean dataset (i.e., tokenized and stemmed) with semantic replacements of emojis to enrich the tweet’s sentiment and enhance the accuracy of distant supervision. The Unicode of an emoji is searched, and the literal description is retrieved. From the literal description, words with emotional orientation are searched, such as ‘love’, ‘happy’, ‘smile’...etc. Then, the emoji is replaced with its equivalent emotional literal. Then, we reflect the annotation on a more natural-looking dataset without sentiment replacement to generate more representative sentence embedding in the next module.

### 3.1.3 Annotation-based on distant supervision

At this stage, the clean text is annotated based on the NRC emotion lexicon (Mohammad & Turney 2013) and Formula (1) following the Distant Supervision Algorithm illustrated in Fig. 2. The proposed lexicon-based annotation approach

$Emotional - Score(T)$ , is calculated accordingly, indicating the emotional orientation of the full tweet’s text. As such, a tweet,  $T$ , is made up of unigrams,  $w$ , such that  $T = \{w_1, w_2, \dots, w_n\}$ . Each unigram may assume a negative or a positive emotional score or may be associated with a distinct emotion such as sadness, anger, happiness, trust, etc., as defined in the NRC lexicon. To account for all negative wellbeing statuses, we aggregate the scores of ‘disgust’, ‘fear’, ‘sadness’, and ‘anger’ with the ‘negative’ emotion score. In addition, we aggregate the scores of ‘happiness’, ‘love’, and ‘trust’ with the ‘positive’ emotional score to account for the positive wellbeing status as explained in Eq. (1). Finally, if the sum of the negative emotional score is greater in a particular tweet than the sum of positive emotional scores, a negative label is assigned, and vice versa. Snowball stemming is applied to ensure the matching of all morphological variations of n-grams with related emotional terms in the lexicon. Neutral labels are assigned to tweets with equal emotional scores. The two procedures accomplishing these steps: *Lexicon\_Based\_Distant\_Supervision(D)* and *Scoring\_Based\_On\_Emotioal\_Lexicon(T)* in Algorithm-1- are illustrated in Fig. 2. Given that Previous studies indicated that emojis could be used as wellbeing indicators conveying various emotional states (Alvarez-Jimenez et al. 2019), emojis’ are replaced with their emotional equivalent based on an emotion replacement matrix, *ER-HashMap*. *ER-HashMap* maps the emoji’s Unicode,  $ew$ , to its corresponding CLDR literal description,  $el$ . This task is performed by the *ReplaceEmoji(ew)* procedure in Algorithm-1- illustrated in Fig. 2. Such that if the emoji’s literal description,  $el$ , contains anything related to the ten explicit emotional terms as defined in the *Common-Emotions-Array*, line 32 in Algorithm-1-, then the explicit emotional term is used instead; otherwise, the literal description of the emoji is retrieved to be evaluated using the NRC lexicon. The proposed annotation algorithm is described in Fig. 2.

$$Emotional - Score(T) =$$

$$\begin{cases} -1 & \text{if } \sum_{i=1}^n |E^+(w_i) + E^{happiness}(w_i) + E^{love}(w_i) + E^{trust}(w_i)| < \sum_{i=1}^n |E^-(w_i) + E^{disgust}(w_i) + E^{fear}(w_i) + E^{sadness}(w_i) + E^{anger}(w_i)| \\ 1 & \text{if } \sum_{i=1}^n |E^+(w_i) + E^{happiness}(w_i) + E^{love}(w_i) + E^{trust}(w_i)| > \sum_{i=1}^n |E^-(w_i) + E^{disgust}(w_i) + E^{fear}(w_i) + E^{sadness}(w_i) + E^{anger}(w_i)| \\ 0 & \text{if } \sum_{i=1}^n |E^+(w_i) + E^{happiness}(w_i) + E^{love}(w_i) + E^{trust}(w_i)| = \sum_{i=1}^n |E^-(w_i) + E^{disgust}(w_i) + E^{fear}(w_i) + E^{sadness}(w_i) + E^{anger}(w_i)| \end{cases} \quad (1)$$

is different from previous methods as it focuses specifically on emotional terms which are considered representative of human mental wellbeing. The proposed approach does not consider only negative or positive words; rather, it focuses on the emotional orientation of the tweet. Therefore, terms with emotional orientation, as defined in the NRC lexicon, are searched in the tweet, and an emotional score,

### 3.2 Module 2: assessment and prediction

Several studies proved that people experiencing wellbeing issues tend to use negative language (De Choudhury et al.

**Algorithm 1** Distant Supervision for Wellbeing Training Dataset

```

1. Inputs:
   •  $D := \text{Twitter Dataset}, D = \{T_1, \dots, T_n\}$ ,  $T: \text{clean and preprocessed Tweet}, T = \{w_1, w_2, \dots, w_l\}$ ,  $w: \text{Unigram Token}$ 
Output:
   •  $ATD := \text{Annotated Twitter Dataset}, ATD = \{(T_j, \text{Emotional Score } (T_j)), \text{Emotional Score } (T) = \{-1, 0, 1\}\}$ 
2. Procedure LEXICON_BASED_DISTANT_SUPERVISION(D)
3. for each Tweet,  $T_j = (\bar{T}_1, \bar{T}_2, \dots, \bar{T}_n) \in D$  do
4.   for each Unigram,  $w_i = (w_1, w_2, \dots, w_l) \in T_j$  do
5.     if  $(w_i \in \{U+0000 \text{ to } U+10FFFF\})$  then,
6.       |  $w_i \leftarrow \text{Replace_Emoji}(w_i)$ 
7.     end if
8.   end for
9.    $\text{Emotional Score } (T_j) \leftarrow \text{Scoring_Based_on_Emotion_Lexicon } (T_j)$ 
10.   $ATD.append(T_j, \text{Emotional Score } (T_j))$ 
11. end for
12. return ATD

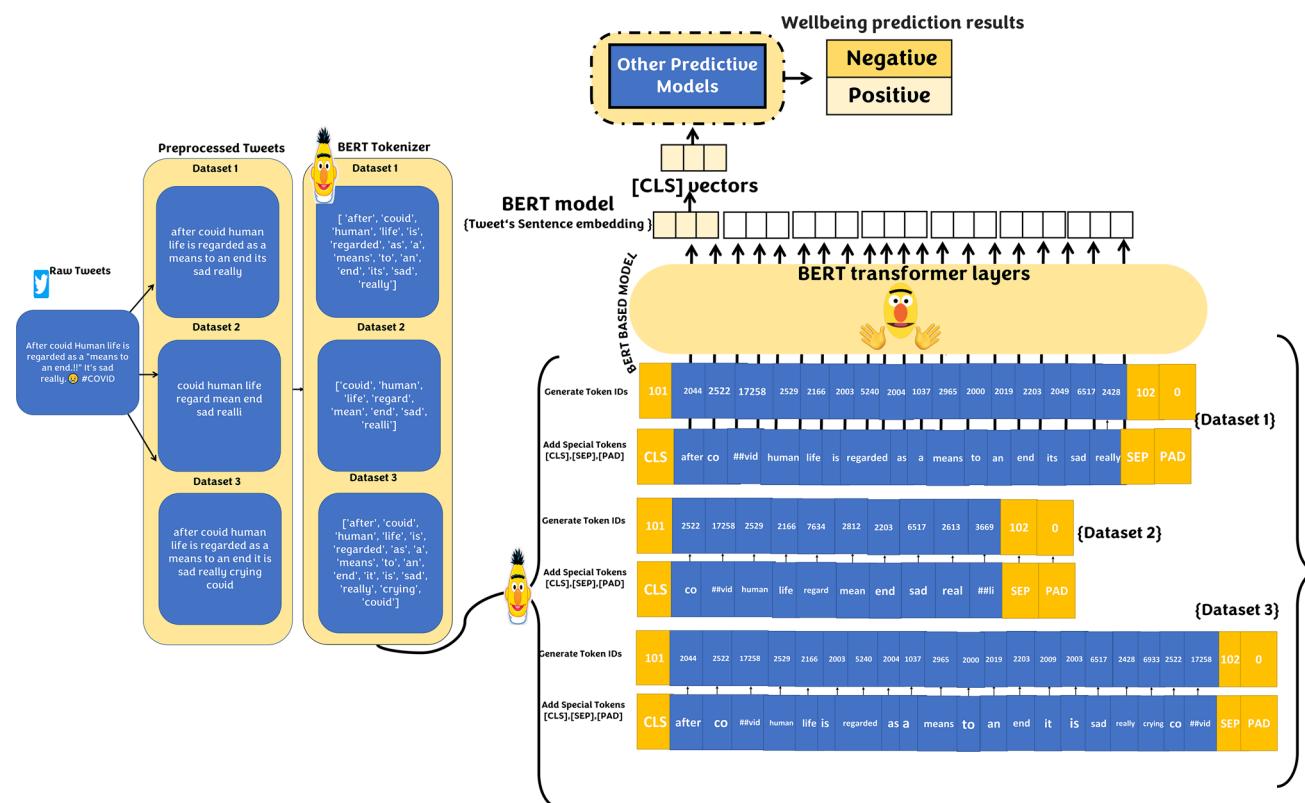
Inputs:
   •  $T := \text{Tweet}, T = \{w_1, \dots, w_l\}$ 
Output:
   •  $\text{Emotional - Score } (T) := \text{Tweet's Emotional Score}, \text{Emotional - Score } (T) = \{-1, 0, 1\}$ 
13. Procedure SCORING_BASED_ON_EMOTIONAL_LEXICON(T)
14.  $NRC := \text{NRC Emotion Lexicon}, NRC = \{(w_i, E(w_i)): i >= 1 \text{ to } |NRC|, w_i \in T_n, E(w_i) \in \{E^+(w_i), E^{happiness}(w_i), E^{love}(w_i), E^{trust}(w_i), E^-(w_i), E^{disgust}(w_i), E^{fear}(w_i), E^{saddness}(w_i), E^{anger}(w_i)\} \}$ 
15. for each Unigram,  $w_i = (w_1, w_2, \dots, w_l) \in T$  do
16.   if  $(E(w_i) \in \{E^+(w_i), E^{happiness}(w_i), E^{love}(w_i), E^{trust}(w_i)\})$  then,
17.     |  $\text{Emotional-Score } ^+(w_i) \leftarrow \text{Emotional-Score } ^+(w_i) + 1$ 
18.   else if  $(E(w_i) \in \{E^-(w_i), E^{disgust}(w_i), E^{fear}(w_i), E^{saddness}(w_i), E^{anger}(w_i)\})$  then,
19.     |  $\text{Emotional-Score } ^-(w_i) \leftarrow \text{Emotional-Score } ^-(w_i) + 1$ 
20.   end if
21. end for
22. if  $(\text{Emotional-Score } ^+(w) > \text{Emotional-Score } ^-(w))$  then,
23.   |  $\text{Emotional - Score } (T) \leftarrow 1$ 
24. else if  $(\text{Emotional-Score } ^+(w) < \text{Emotional-Score } ^-(w))$  then,
25.   |  $\text{Emotional - Score } (T) \leftarrow -1$ 
26. else
27.   |  $\text{Emotional - Score } (T) \leftarrow 0$ 
28. end if
29. return  $\text{Emotional - Score } (T)$ 

Inputs:
   •  $ew := \text{Emoji Unigram Unicode}, ew = \{U+0000 \text{ to } U+10FFFF\}$ 
Output:
   •  $lw := \text{Literal Unigram}$ 
30. Procedure REPLCAE_EMOJI(ew)
31.  $lw \leftarrow \text{ER-Matrix } (ew) // \text{s.t. ER-HashMap} = \{(ew, lw)\}$ ,  $ew: \text{Emoji Unigram Unicode}, ew \in \{U+0000 \text{ to } U+10FFFF\}$ ,  $lw: \text{Emoji Description} := \text{CLDR Short Name}$ 
32. Common-Emotions-Array [] = ["love", "happy", "sad", "smile", "angry", "gring", "cry", "anxious", "confound", "scream"]
33. for  $j$  from 0 to  $| \text{Common-Emotions-Array} - 1 |$ 
34.   if  $\text{stem}(lw) = \text{stem}(\text{Common-Emotions-Array}[j])$ ,  $lw \leftarrow \text{Common-Emotions-Array}[j]$ 
35. else  $lw \leftarrow lw$ 
36. return  $lw$ 
```

**Fig. 2** Distant supervision algorithm for wellbeing dataset

2013; Park et al. 2015; Tsugawa et al. 2015); hence, we aim to evaluate public wellbeing through the use of negative language. In this module, first, real-time Twitter chatter data is retrieved and cleaned using similar pre-processing tasks explained in Module-1-. Then, the BERT model is used to generate rich contextual sentence embedding for the new chatter data and the labelled training data generated

in Module-1-. Second, using the labelled sentence embeddings, several classifiers are trained to predict, in real-time, whether a user is experiencing wellbeing challenges or not. We use BERT<sub>BASE</sub> ( $L = 12, H = 768, A = 12$ , Total Parameters = 110 M), where L is the number of layers, the hidden size as H, and the number of self-attention heads as A. We consider a tweet as a sentence, i.e., a sequence of tokens



**Fig. 3** BERT-based context-specific sentence embedding

composed of one or multiple sentences. More formally, let us consider a tweet as  $X = \{x_1, \dots, x_n\}$ , where  $n$  denotes the length of the Tweet such that  $|X| = n$ . BERT adds two special tokens to the input sequence. The first token of every sequence is always a special classification token ([CLS]), and the last token is a separation token ([SEP]). Hence, in our problem, a Tweet is defined as  $X = \{\text{CLS}, x_1, \dots, x_n, \text{SEP}\}$ . BERT uses the final hidden state corresponding to ([CLS]) token as the aggregate sequence representation for classification tasks. Therefore, the output corresponding to ([CLS]) token can be thought of as an embedding for the entire sentence (i.e., tweet). We obtain the sentence embedding for all the tweets in the labelled training dataset as well as in the real-time chatter data by extracting the final vector output of this token. The labelled sentence embeddings are used as the feature set to train several classifiers. After that, the learned model is used to predict the wellbeing indicators in the new unlabelled real-time Twitter chatter data. Figure 3 illustrates how context-specific sentence embeddings for the tweets are generated using BERT model for the three experimental datasets listed in Sect. 4.2.1. As can be seen in Figure 3, for tweets' (i.e., sentences') wellbeing classification, we're only interested in the first vector of the BERT's output associated with the [CLS] token. That is because BERT is trained to encapsulate a sentence-wide sense of the output at the

first position (i.e., at the CLS token) producing a contextual sentence embedding of the input. These contextual sentence embeddings are then used to train other predictive models to produce wellbeing predictions.

### 3.3 Module 3: analytics and visualization

This module aims to assess and visualize the variability in the emotional status of the public through their Tweets over time across several locations based on the predictions generated in Module#2 and the associated meta-data. After conducting spatiotemporal analysis, and interaction analysis, we can measure the variability of public wellbeing indicators at different times of the day, over several time events, and across several cities aiding decision-makers in planning necessary intervention plans. For instance, the public reaction to schools' closure might differ in emotional intensity from one city to another. By analyzing the variations in the intensity of public emotion relevant to different temporal events in several locations, the decision-makers can make more well-informed decisions. In addition, analyzing public emotions over several time intervals in the day provides crucial evidence of public wellbeing. Statistics show that nighttime online activity is a known characteristic of individuals susceptible to wellbeing

**Table 2** Tweets before and after preprocessing and emoji replacement

| Original Tweet   | Processed tweet  | Annotation |
|--|--|------------|
| We can heal the Earth ❤️ <a href="https://t.co/RwrL0BR1Vf">https://t.co/RwrL0BR1Vf</a>   | we can heal the earth love   | Positive   |
| Please retweet. Yes Trump said worries about cornov is the new hoax 🤦. <a href="https://t.co/SsKL9Wuq4L">https://t.co/SsKL9Wuq4L</a> | please retweet yes trump said worries about cornov is the new hoax angry | Negative   |

issues, as highlighted and validated in (De Choudhury et al. 2013; Lustberg & Reynolds 2000). Hence, an increased nighttime Twitter activity can indicate the rise of wellbeing issues during the pandemic, besides other linguistic indicators.

Moreover, interaction analysis measures users' responses to negative tweets after analyzing "retweet" and "favorite" activities. It is prevalent on social media platforms to have people navigating and surfing but never posting content. The interaction of this group of the public can give insight into their wellbeing status. As such, people interacting with negative Tweets are more likely to be experiencing negative emotions and vice versa. To perform temporal and spatial analysis, metadata is retrieved from Twitter JSON object indicating cities in the country, interactions in the form of "Retweets" or "Likes", as well as the time-related attributes. Therefore, several spatiotemporal and linguistic visualizations, such as sentiment maps, wordclouds, and sentiment bars, are used to provide insight into the public's emotional status.

## 4 Experimental work and results

In this section, we demonstrate experimental results as follows:

Section 4.1 validates the proposed distant supervision approach to generate a wellbeing-related training dataset from unstructured text.

Section 4.2 validates the proposed BERT-based wellbeing prediction model.

Section 4.3 demonstrates how wellbeing predictions from Twitter can reveal decision-support critical insights through a case study related to the education sector.

### 4.1 Validating distant supervision method

To validate the proposed distant supervision method for wellbeing dataset labeling, we use the first published COVID-19 Twitter dataset, collected on January 22, 2020 (Chen et al. 2020). This data is selected as it contains intensive wellbeing indicators considering the context of the COVID-19 outbreak. Twitter chatter data is preprocessed as explained in 3.1.1 and 3.1.2, and an overall emotional score of the tweet is calculated as explained in Fig. 2. Table 2 shows examples of tweets before and after pre-processing.

Next, manual validation is conducted on two annotated datasets by human expert annotators. Each dataset contains 16,797 tweets. Using stratified sampling, two representative samples of 580 tweets are selected from each dataset to ensure balanced examples proportional to each class label with a 95% confidence level and confidence interval of 4. We evaluated the agreement between experts and the method's annotation using Kappa interrater agreement test. Kappa scores were 0.85 and 0.81 for the first and second samples, respectively. These Kappa scores indicate good agreement between the human rater and the proposed annotation method rating. Minor discrepancies between the proposed method's annotation and the expert's annotation were identified due to the extensive use of informal language on Twitter. For example, cases of sarcasm and irony are not perfectly captured by the distant supervision approach, especially for concise tweets. However, those were minor cases as typically, users would use irony in conjunction with some representative emojis which are handled in our approach. In addition, negation is not captured in distant supervision; the method does not consider "I am not happy" as a negative sentence because it contains the word 'happy' which has a positive emotional sentiment. However, in psychology, a person saying that they are "happy" or "not happy" is still demonstrating a focus on positivity as a dimension of thought. They could have just as easily said "sad" or "angry", but it is psychologically telling that they are thinking in terms of "happiness" (Pennebaker et al. 2003; Tausczik & Pennebaker 2010). This, in turn, supports the validity of the proposed distant supervision approach despite the few discrepancies caused by negation or sarcasm.

### 4.2 Validating wellbeing predictions with BERT-based sentence embeddings

In this section, we validate the proposed approach for wellbeing predictions using BERT as a language model to provide rich contextualized sentence embeddings in conjunction with several classifiers. The purpose of these experiments is to prove that BERT as a language model gives more accurate

**Table 3** Statistics of evaluation dataset

| Data            | Size   |
|-----------------|--------|
| Total           | 10,888 |
| Negative tweets | 9369   |
| Positive tweets | 1519   |

predictions, to select the best classifier to be used for online predictions with Twitter data, as well as to select the best pre-processing tasks that would generate the highest prediction results.

Using distant supervision, explained in the previous section, an evaluation wellbeing dataset is generated. We consider original Tweets only, i.e., retweets and Quoted Tweets are omitted from this dataset. Also, repeated Tweets are removed. Hence the evaluation dataset contains unique Tweets collected during March 2020 upon declaring COVID-19 as a world pandemic as it is expected to be rich in wellbeing indicators. Statistics of the evaluation dataset after labelling using the proposed distant supervision method are summarized in Table 3. We also omit neutral Tweets as we are focusing on the positive and negative dimensions of thoughts.

#### 4.2.1 Pre-processing of the evaluation dataset

To identify the best pre-processing required for the assessment module, three versions of the evaluation dataset are generated; each had undergone different pre-processing tasks, as follows:

1. Dataset-1-: A dataset cleaned from URLs, mentions, hashtags, emoji, and special characters without any constraint on character length and stopwords are kept, i.e., Tweets are kept as natural sentences without abnormal tokens.
2. Dataset-2-: A dataset with similar pre-processing applied to Dataset-1; in addition, stopwords are removed, and the minimum character length is set to two characters. This is the typical pre-processing done in most of the sentiment and emotion analysis tasks (Ismail et al. 2018).
3. Dataset-3-: A dataset with similar pre-processing applied to Dataset-1, except that emojis are kept and replaced with their emotional equivalent, as illustrated in Table 2. Besides, abbreviated verbs are transformed into full form. Also, hashtags are kept, but the '#' special character is removed. This is done to ensure that all emotions conveyed by a Twitter user are kept in the text. A task that is commonly done in emotion analysis.

#### 4.2.2 Baselines with different feature vector models

To compare the performance of the classification model trained on sentence embeddings using BERT with other feature representations, we generate three baselines with different feature vector representations, commonly used in the literature on wellbeing automatic assessment (Husseini Orabi et al. 2018; Resnik et al. 2015; Tsugawa et al. 2015), as follows:

**Table 4** Classification results for dataset-1

| Features | Measures  | LR   | KNN  | SVM  | DT   |
|----------|-----------|------|------|------|------|
| BERT     | Precision | 0.81 | 0.86 | 0.87 | 0.84 |
|          | F-measure | 0.85 | 0.87 | 0.88 | 0.84 |
| BoW      | Precision | 0.7  | 0.67 | 0.87 | 0.61 |
|          | F-measure | 0.47 | 0.56 | 0.54 | 0.61 |
| NRC      | Precision | 0.45 | 0.47 | 0.45 | 0.47 |
|          | F-measure | 0.47 | 0.47 | 0.47 | 0.47 |
| Word2Vec | Precision | 0.64 | 0.71 | 0.63 | 0.60 |
|          | F-measure | 0.62 | 0.54 | 0.62 | 0.60 |

**Table 5** Classification results for dataset-2

| Features | Measures  | LR   | KNN  | SVM  | DT   |
|----------|-----------|------|------|------|------|
| BERT     | Precision | 0.74 | 0.83 | 0.84 | 0.8  |
|          | F-measure | 0.8  | 0.83 | 0.85 | 0.8  |
| BoW      | Precision | 0.68 | 0.63 | 0.80 | 0.58 |
|          | F-measure | 0.46 | 0.57 | 0.56 | 0.58 |
| NRC      | Precision | 0.43 | 0.56 | 0.43 | 0.53 |
|          | F-measure | 0.46 | 0.51 | 0.46 | 0.47 |
| Word2Vec | Precision | 0.7  | 0.81 | 0.69 | 0.60 |
|          | F-measure | 0.69 | 0.52 | 0.68 | 0.60 |

1. Baseline-1-: Term frequency inverse document frequency (TF-IDF) BoW features.
2. Baseline-2-: NRC emotion lexicon features are populated with frequencies.
3. Baseline-3-: Word2Vec, word embeddings' features.

#### 4.2.3 Experiments and results

For wellbeing's prediction evaluation, support vector machines (SVM), K-nearest Neighbors (KNN), Logistic Regression (LR), and Decision Trees (DT), classifiers are trained on the full dataset using the different variants of the evaluation dataset explained earlier (i.e., Dataset-1-, Dataset-2-, Dataset-3-) with three different feature vector models. The predictive models are evaluated with stratified twofold cross-validation.

As illustrated in Table 3, the class distribution of the evaluation dataset is considerably skewed towards the negative class, which is caused by the increased negative feelings evolving post the COVID-19 outbreak during which the evaluation data was collected. This imbalanced class distribution is known to generate inaccurate classification results during evaluation and can lead to model overfitting problems during the training phase (Branco et al. 2015; Galar et al. 2012; Powers 2015). Hence, stratified cross-validation is used to avoid overfitting and ensure each class

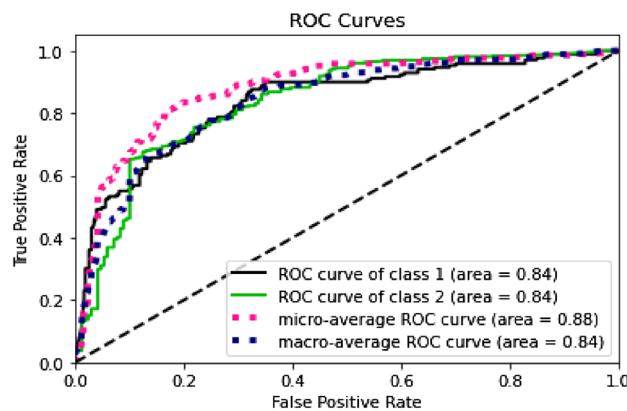
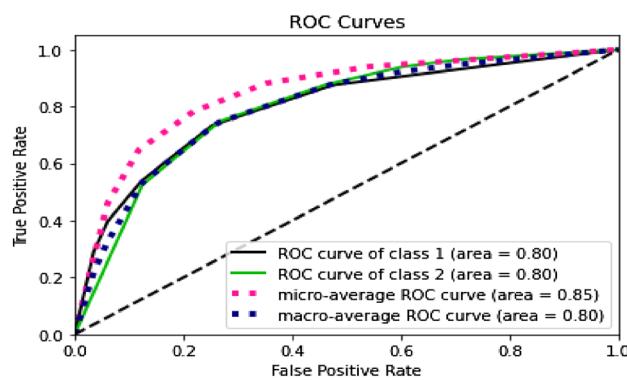
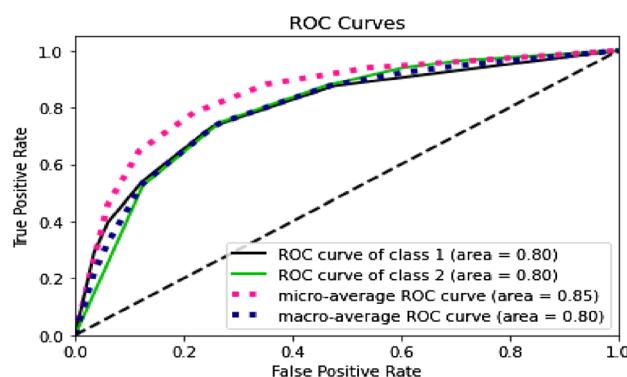
**Table 6** Classification results for dataset-3

| Features | Measures  | LR   | KNN  | SVM  | DT   |
|----------|-----------|------|------|------|------|
| BERT     | Precision | 0.81 | 0.87 | 0.87 | 0.85 |
|          | F-measure | 0.85 | 0.87 | 0.88 | 0.84 |
| BoW      | Precision | 0.7  | 0.67 | 0.87 | 0.61 |
|          | F-measure | 0.47 | 0.56 | 0.53 | 0.61 |
| NRC      | Precision | 0.45 | 0.45 | 0.45 | 0.47 |
|          | F-measure | 0.47 | 0.47 | 0.47 | 0.47 |
| Word2Vec | Precision | 0.64 | 0.78 | 0.64 | 0.57 |
|          | F-measure | 0.62 | 0.52 | 0.62 | 0.57 |

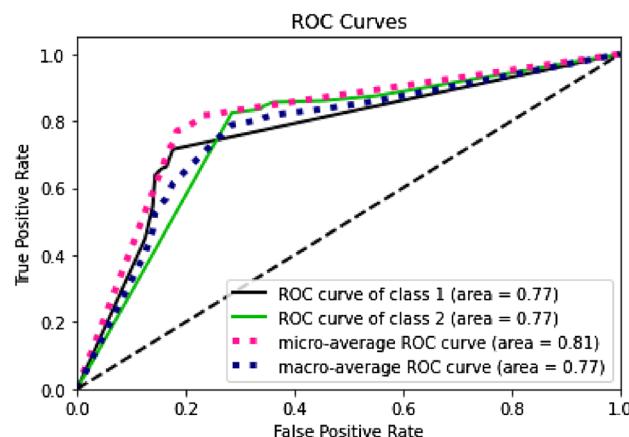
is approximately equally represented across each test fold proportionally to class distribution in the full dataset.

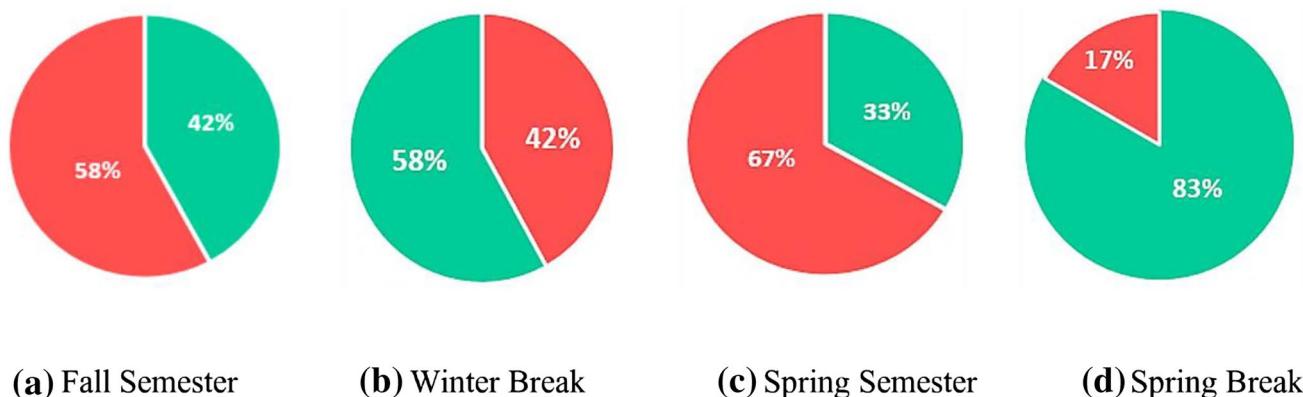
In addition, to accurately evaluate the quality for classification despite class skewness, precision and F-measures are selected to evaluate the classification quality (Liu et al. 2009). Precision indicates how well a classification model identifies positive classes (i.e., the majority class, and in our dataset, it represents tweets with negative emotions). We consider this as an essential quality criterion in the selected classification model. That is because we aim to specifically identify users susceptible to experiencing wellbeing challenges. On the other hand, the F-measure, i.e., the weighted average of Precision and Recall, is used to evaluate the classification balance between specificity and completeness. The higher the F-measure in this problem, the more balance between positive cases detected and the total number of positive cases in the dataset.

Looking at the classification results in Tables 4, 5 and 6, we can see that sentence embedding features, based on BERT, consistently generate highly precise results (between 0.81 and 0.87) using multiple classifiers trained on dataset-1-, i.e., more natural-looking text, and dataset-3-, i.e., with emotion replacement, as opposed to all other feature models. In contrast, feature vector models commonly used

**Fig. 5** SVM**Fig. 6** KNN**Fig. 7** LR

in the literature of wellbeing assessment, i.e., BoW and emotion features, generate lower precision values using all the evaluation datasets. This, in turn, supports using BERT for contextualized feature representation, especially with the possibility of using less processed text to get highly precise results. As such, this would work well with real-time data

**Fig. 4** Decision tree



**Fig. 8** Overall sentiment score across the UAE from fall semester 2020 to spring break 2021

as we are in favor of reducing pre-processing overhead and predicting public wellbeing from the raw text as much as possible. BERT-based features also generate a good balance between positively detected cases out of all positive cases in the dataset as depicted by the F-measure (between 0.84 and 0.88) using dataset-1- and dataset-3- as opposed to all other feature vector models.

Interestingly, classification quality slightly drops when training classifiers with BERT-based features using dataset-2-, i.e., without stopwords and short tokens. Precision reaches its minimum value of 0.74 with LR when using dataset-2- with BERT. This is because BERT is trained on natural text, including stopwords, which makes it more powerful in understanding the full context. Even though word embeddings are known to generate rich semantics, i.e., adding context to words, word2vec feature vectors generated very low precision values (between 0.57 and 0.81) when compared with BERT's.

Furthermore, by looking at the ROC curves in Figs. 4, 5, 6 and 7, we can see that the prediction models trained on BERT sentence embeddings as feature vectors based on Dataset-3-had consistently generated high accuracies with AUC ranging between 0.77 and 0.85 with the four classifiers (i.e., DT, SVM, KNN, and LR) for both classes. This in turn surpasses all reported prediction results in the literature of automatic wellbeing assessment based on Twitter dataset, as summarized in Table 1. We can further see that the prediction accuracy for both classes (i.e., the negative class and positive class) are, to some extent, comparable despite the imbalanced distribution of class labels. In addition, considering the micro average of the ROC Curve for both classes, the AUC ranges between 0.81 to 0.89 with the four classifiers (i.e., DT, SVM, KNN, and LR), which gives highly accurate predictions for the majority class label, which in this case the class label related to negative emotions. Also, when comparing the results of macro average

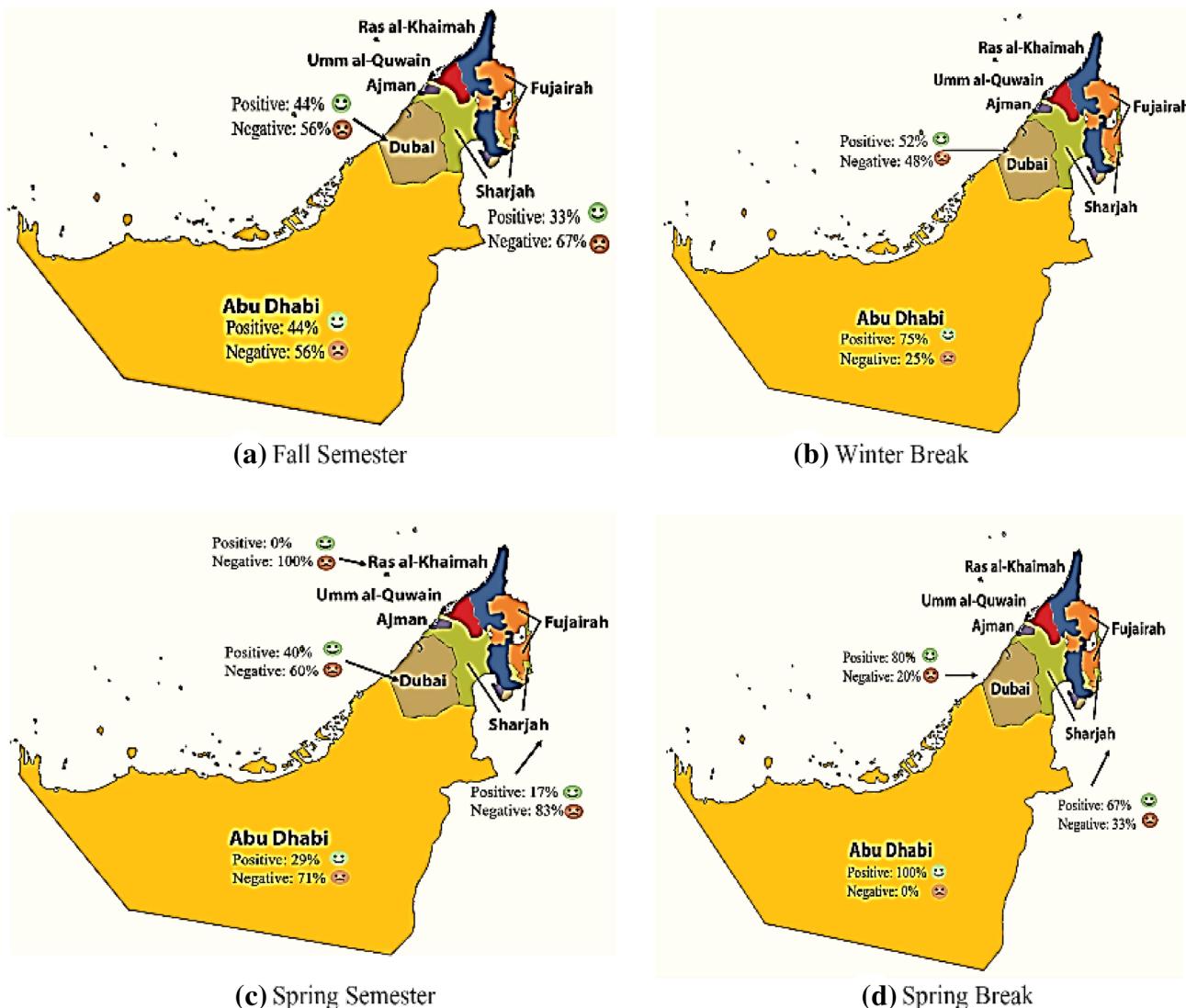
and micro average for both class labels, the difference is not significant, indicating that prediction accuracies are not significantly impacted by the skewness of class distribution and are not biased to a particular class label. This suggests that the proposed wellbeing prediction model trained on BERT and labelled data using the proposed distant supervision method can give accurate wellbeing predictions for Twitter chatter data.

#### 4.3 Evaluating the utility of the proposed framework in evaluating public wellbeing

To validate the utility of the proposed method in assessing public wellbeing in a particular context, a UAE-specific dataset was generated and targeted around COVID-19 to ensure the richness of wellbeing indicators. We generated the first bilingual (Arabic and English) geotagged, COVID-19 specific dataset that is focused on tweets generated from various locations in the UAE. The evaluation Twitter Chatter data was collected on an hourly basis using the Twitter API since October 10th, 2020, by tracking keywords related to the novel coronavirus.<sup>1</sup> The English keywords used to filter the tweets are *Epidemic, Pandemic, Corona, COVID, Sterilization, Quarantine, Wuhan, Coronavirus*.

We used the proposed framework to predict wellbeing status from the public Twitter chatter data relevant to a specific sector, namely, Education. The goal is to demonstrate how online learning has affected the wellbeing of students, teachers, and parents during the pandemic and how we can support decision-makers with information to make appropriate decisions. It is apparent that the pandemic has caused disruption in all educational institutions around the world. For evaluation and illustration purposes, four timeframes were selected: (a) School Fall Semester, (b) School Winter

<sup>1</sup> <https://github.com/RamiAltai/TwitterData>



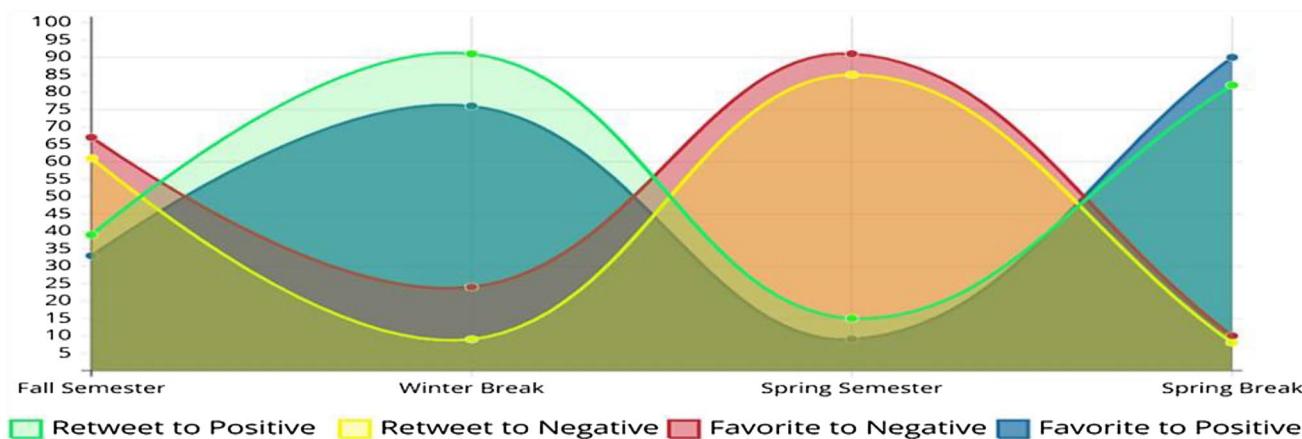
**Fig. 9** Spatial emotion distribution across the UAE

break, (c) School Spring Semester, and (d) School Spring Break.

Figure 8 illustrates the positive to negative overall scores for each of the defined timeframes in pie charts. The red colour indicates negative emotions, while the green color indicates positive emotions. Figure 8a and c show a negative majority in the analyzed tweets, which happen to be during the first two semesters of the educational institutions' closure and the migration to e-learning (i.e., Fall Semester 2020 and Spring Semester 2020). On the other hand, Fig. 8b and d show a positive majority during the breaks. Moreover, we can see an increased negativity score in the Spring semester compared to the Fall semester. In the fall semester, some schools managed to support a hybrid delivery mode combining face-to-face and online learning, whereas, in the Spring semester, a full lockdown was forced across the

country, except in Dubai, resulting in increased overall negative emotions.

Figure 9 provides a visualization of the positive and negative emotion distribution across the different emirates of the UAE. Map visualization provides spatial analytics of emotions across several emirates. In Fig. 9, we can see the change of emotion across the emirates of Abu Dhabi, Dubai, Sharjah, and Ras Al Khaimah during the four defined timeframes. It is obvious that negativity had increased significantly in all the emirates but specifically in the emirates of Abu Dhabi, Sharjah, and Ras Al Khaimah, upon the compulsory full school closure in the Spring semester. Dubai, however, managed to continue offering full face-to-face study in addition to the hybrid mode of study, which can be seen in the minor drop in positivity compared with the other emirates.



**Fig. 10** Reactions to positive versus negative tweets

In addition to the above statistics, the proposed model analyzes the interaction to positive and negative tweets given that many people might not always play an active role on social media accounts and post tweets; rather, they would just react to posted tweets. Reactions indicate people's emotions and opinions as well. Figure 10 illustrates these reactions and their change across the four defined timeframes on both positive and negative tweets. By analyzing people's reactions, we can see that the public's wellbeing indicators are consistent across the four timeframes defined for the study showing increased positivity during breaks and increased negativity during semesters with increased reactions to negative tweets during the compulsory lockdown period During Semester 2020.

As illustrated above, social media can give insights into the public wellbeing status which can support decision-makers in the decision-making process. Using this framework, governments can gauge public opinion towards specific events. For instance, by defining specific temporal events such as the first introduction of a new online examination framework', or the rectification of new rules in the education sector, officials can gauge the public's reaction to these new decisions and whether a change is required.

## 5 Conclusion

The public wellbeing situation is directly and significantly impacted by government decisions around the globe. With the ever-changing situations and pace of modern life, public wellbeing is prioritized. Health authorities find themselves obliged to analyze the wellbeing of the public and have discovered that social media platforms are very rich sources of data that are full of emotions and wellbeing indicators. Therefore, a solution that relies on analyzing massive and

spatiotemporal social media data with a focus on wellbeing will provide impactful insights.

In this paper, an intelligent decision support system is proposed encompassing several modules. The proposed system infers spatiotemporal insights on the public's wellbeing from social media content through the implementation of a novel distant supervision approach as well as an emotion prediction model trained on semantically rich sentence embedding using BERT. Several experiments are conducted to validate the effectiveness of the proposed distant supervision approach, the prediction model, as well as the usefulness of the generated insights in supporting decision making. A further extension to the proposed framework would consider analyzing emotional triggers causing several negative and positive emotions in different sectors.

**Author contributions** H.I. Conceived the idea, proposed the analytics framework and the distant supervision approach, contributed to development, contributed to writing all sections of the manuscript, and supervised the overall work, M.S. contributed to supervision and to writing the case study section, N.H. Contributed to writing, modeling, analysis, and data curation, R.E. contributed to modeling and analysis, A.N. contributed to development.

**Funding** This research project was funded by the Office of Research and Sponsored Programs, Abu Dhabi University and Zayed Health Center, UAE University (Fund#12R005)

## Declarations

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

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated

otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Allen J, Cotter-Roberts A, Kadel R, Hughes K, Dyakova M (2021) COVID-19 impact on financial security: evidence from the National Public Engagement Survey in Wales. *Eur J Public Health.* <https://doi.org/10.1093/eurpub/ckab165.324>
- Almouzini S, Khemakhem M, Alageel A (2019) Detecting Arabic depressed users from twitter data. *Procedia Comput Sci* 163:257–265. <https://doi.org/10.1016/j.procs.2019.12.107>
- Alvarez-Jimenez M, Murray G, McLeod H, Allan S, Van Dam L, Rietstra S, Van der Drift E, Jan M, Stams GJ, Van der Mei R, Mahfoud M, Popma A, Schlossberg E, Pentland A, Reid TG (2019) Can an emoji a day keep the doctor away? An explorative mixed-methods feasibility study to develop a self-help app for youth with mental health problems. *Psychiatry* 10:593. <https://doi.org/10.3389/fpsyg.2019.00593>
- Boon-Itt S, Skunkan Y (2020) Public perception of the COVID-19 pandemic on twitter: sentiment analysis and topic modeling study. *JMIR Public Health Surveill* 6(4):e21978. <https://doi.org/10.2196/21978>
- Branco P, Torgo L, & Ribeiro R (2015) A survey of predictive modeling under imbalanced distributions
- Bustos VP, Comer CD, Manstein SM, Laikhter E, Shiah E, Xun H, Lee BT, Lin SJ (2022) Twitter voices: twitter users' sentiments and emotions about COVID-19 vaccination within the United States. *Eur J Environ Public Health* 6(1):em0096. <https://doi.org/10.21601/EJEPH/11499>
- Chatterjee A, Narahari KN, Joshi M, & Agrawal P (2019) SemEval-2019 task 3: EmoContext contextual emotion detection in text. In: NAACL HLT 2019 - international workshop on semantic evaluation, SemEval 2019, proceedings of the 13th workshop, pp. 39–48. Doi: <https://doi.org/10.18653/V1/S19-2005>
- Chen E, Lerman K, Ferrara E (2020) Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus twitter data set. *JMIR Public Health Surveill.* <https://doi.org/10.2196/19273>
- Chintalapudi N, Battineni G, Amenta F (2021) Sentimental analysis of COVID-19 tweets using deep learning models. *Infect Dis Rep* 13(2):329. <https://doi.org/10.3390/IDR13020032>
- De Choudhury M, Gamon M, Counts S, & Horvitz E (2013) Predicting depression via social media. In: Proceedings of the 7th international conference on weblogs and social media, ICWSM 2013, pp. 128–137
- Coppersmith G, Dredze M, & Harman C (2015) Quantifying mental health signals in twitter. pp. 51–60. Doi: <https://doi.org/10.3115/W14-3207>
- Devlin J, Chang M-W, Lee K, Google KT, & Language AI (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the north, pp. 4171–4186. doi: <https://doi.org/10.18653/V1/N19-1423>
- Dubey AD (2020) Twitter sentiment analysis during COVID-19 outbreak. *SSRN Electron J.* <https://doi.org/10.2139/SSRN.3572023>
- Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans Syst Man Cybern Part C Appl Rev* 42(4):463–484. <https://doi.org/10.1109/TSMCC.2011.2161285>
- Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC (2017) Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci* 18:43–49. <https://doi.org/10.1016/j.cobeha.2017.07.005>
- Orabi AH, Buddhitha P, Orabi MH, & Inkpen D (2018) Deep learning for depression detection of twitter users. *Aclweb.Org*, pp. 88–97. Doi: <https://doi.org/10.18653/v1/w18-0609>
- Inui T, Umesawa Y, Yamamoto M (2013) Complaint sentence detection via automatic training data generation using sentiment lexicons and context coherence. *J Nat Lang Process* 20(5):683–705. <https://doi.org/10.5715/jnlp.20.683>
- Ismail HM, Belkhouche B, Zaki N (2018) Semantic Twitter sentiment analysis based on a fuzzy thesaurus. *Soft Comput* 22(18):6011–6024. <https://doi.org/10.1007/s00500-017-2994-8>
- Jones S, Murphy F, Edwards M, James J (2008) Doing things differently: advantages and disadvantages of Web questionnaires. *Nurse Res* 15(4):15–26. <https://doi.org/10.7748/NR2008.07.15.4.15.C6658>
- Khaledi M, Haghhighatdoost F, Feizi A, Aminorroaya A (2019) The prevalence of comorbid depression in patients with type 2 diabetes: an updated systematic review and meta-analysis on huge number of observational studies. *Acta Diabetologica* 56(6):631–650. <https://doi.org/10.1007/S00592-019-01295-9>
- Kleinberg B, Vegter I van der, & Mozes M (2020) Measuring emotions in the COVID-19 real world worry dataset. <https://aclanthology.org/2020.nlp covid19-acl.11>
- Krendl AC, Perry BL (2021) The impact of sheltering in place during the COVID-19 pandemic on older adults' social and mental well-being. *J Gerontol Ser B* 76(2):e53–e58. <https://doi.org/10.1093/GERONB/GBAA110>
- Kydros D, Argyropoulou M, Vrana V (2021) A content and sentiment analysis of greek tweets during the pandemic. *Sustainability* 13(11):6150. <https://doi.org/10.3390/SU13116150>
- Li S, Wang Y, Xue J, Zhao N, 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. <https://doi.org/10.3390/ijerph17062032>
- Li X, Zhou M, Wu J, Yuan A, Wu F, & Li J (2020b) Analyzing COVID-19 on online social media: trends, sentiments and emotions
- Liu X-Y, Wu J, Zhou Z-H, Member S (2009) Exploratory undersampling for class-imbalance learning. *Cybernetics* 39(2):539. <https://doi.org/10.1109/TSMCB.2008.2007853>
- Lustberg L, Reynolds CF (2000) Depression and insomnia: questions of cause and effect. *Sleep Med Rev* 4(3):253–262. <https://doi.org/10.1053/smrv.1999.0075>
- Mohammad SM, Turney PD (2013) Crowdsourcing a word-emotion association lexicon. *Comput Intell* 29(3):436–465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- Park G, Andrew Schwartz H, Eichstaedt JC, Kern ML, Kosinski M, Stillwell DJ, Ungar LH, Seligman MEP (2015) Automatic personality assessment through social media language. *J Pers Soc Psychol* 108(6):934–952. <https://doi.org/10.1037/pspp0000020>
- Patrick SW, Henkhaus LE, Zickafoose JS, Lovell K, Halvorson A, Loch S, Letterie M, Davis MM (2020) Well-being of parents and children during the COVID-19 pandemic: a national survey. *Pediatrics* 146(4):e2020016824. <https://doi.org/10.1542/PEDS.2020-016824>
- Pennebaker JW, Mehl MR, Niederhoffer KG (2003) Psychological aspects of natural language use: our words, our selves. *Annu Rev Psychol* 54(1):547–577. <https://doi.org/10.1146/annurev.psych.54.101601.145041>

- Pennebaker JW, Booth RJ, Boyd RL, & Francis ME (n.d.) Linguistic inquiry and word count: LIWC2015 operator's manual. In downloads.liwc.net.s3.amazonaws
- Powers DMW (2015) What the F-measure doesn't measure: features, flaws, fallacies and fixes
- Reece AG, Reagan AJ, Lix KLM, Dodds PS, Danforth CM, Langer EJ (2016) Forecasting the onset and course of mental illness with Twitter data. *Sci Rep* 7(1):1–11
- Resnik P, Garron A, & Resnik R (2013) Using topic modeling to improve prediction of neuroticism and depression in college students. In: EMNLP 2013 - 2013 conference on empirical methods in natural language processing, proceedings of the conference, pp. 1348–1353
- Resnik P, Armstrong W, Claudino L, Nguyen T, Nguyen V-A, & Boyd-Graber J (2015) Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. pp. 99–107. Doi: <https://doi.org/10.3115/v1/w15-1212>
- Samuel J, Nawaz Ali GG, Rahman MM, Esawi E, & Samuel Y (n.d.) COVID-19 public sentiment insights and machine learning for tweets classification. Doi: <https://doi.org/10.3390/info11060314>
- Scheuer C, Boot E, Carse N, Clardy A, Gallagher J, Heck S, Marron S, Martinez-Alvarez L, Masarykova D, Mcmillan P, Murphy F, Steel E, Ekdom H Van, & Vecchione H (2020) Measuring emotions in the COVID-19 real world worry dataset. *Physical Education and Sport for Children and Youth with Special Needs Researches – Best Practices – Situation*, pp. 343–354
- Schwartz H, Eichstaedt J, Kern M, Park G, Sap M, Stillwell D, Kosinski M, & Ungar L (2014) Towards assessing changes in degree of depression through Facebook. In: Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality
- Shahriarirad R, Erfani A, Ranjbar K, Bazrafshan A, Mirahmadizadeh A (2021) The mental health impact of COVID-19 outbreak: a Nationwide Survey in Iran. *Int J Mental Health Syst*. <https://doi.org/10.1186/S13033-021-00445-3>
- Shetty NP, Muniyal B, Anand A, Kumar S, Prabhu S (2020) Predicting depression using deep learning and ensemble algorithms on raw twitter data. *Int J Electr Comput Eng* 10(4):3751–3756. <https://doi.org/10.11591/ijece.v10i4.pp3751-3756>
- Son C, Hegde S, Smith A, Wang X, Sasangohar F (2020) Effects of COVID-19 on college students' mental health in the United States: interview survey study. *J Med Internet Res* 22(9):14. <https://doi.org/10.2196/21279>
- Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol* 29(1):24–54. <https://doi.org/10.1177/0261927X09351676>
- Thapa S, Sotang N, Adhikari J, Ghimire A, Limbu AK, Joshi A, Adhikari S (2020) Impact of COVID-19 lockdown on agriculture education in Nepal: an Online survey. *Pedagog Res*. <https://doi.org/10.2933/pr/8465>
- Tsugawa S, Mogi Y, Kikuchi Y, Kishino F, Fujita K, Itoh Y, Ohsaki H (2013) On estimating depressive tendencies of Twitter users utilizing their tweet data. *Proc IEEE Virtual Real*. <https://doi.org/10.1109/VR.2013.6549431>
- Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, & Ohsaki H (2015). Recognizing depression from twitter activity. In: Conference on human factors in computing systems - proceedings, 2015–April, 3187–3196. Doi: <https://doi.org/10.1145/2702123.2702280>
- VanderWeele TJ, Fulks J, Plake JF, Lee MT (2021) National well-being measures before and during the COVID-19 pandemic in online samples. *J Gen Intern Med* 36(1):248–250. <https://doi.org/10.1007/S11606-020-06274-3/TABLES/2>
- Wang C, Pan R, Wan X, Tan Y, Xu L, Ho CS, Ho RC (2020) Immediate psychological responses and associated factors during the initial stage of the 2019 coronavirus disease (COVID-19) epidemic among the general population in China. *Int J Environ Res Public Health* 17(5):1729. <https://doi.org/10.3390/ijerph17051729>
- Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. *IEEE Comput Intell Mag* 13(3):55–75. <https://doi.org/10.1109/MCI.2018.2840738>
- Zhou J, Liu L, Xue P, Yang X, Tang X (2020) Mental health response to the COVID-19 outbreak in China. *Am J Psychiatry* 177(7):574–575. <https://doi.org/10.1176/Appi.Ajp.2020.20030304>

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