Data Article

Modern Standard Arabic mood changing and depression dataset

Ashwag Maghraby*, Honsia Ali

Computer Science Department, College Of Computer Science and Information, Umm Al-Qura University, Mecca, Saudi Arabia

ARTICLE INFO

Article history:
Received 5 January 2022
Revised 17 February 2022
Accepted 23 February 2022
Available online 1 March 2022

Dataset link: Modern Standard Arabic mood changing and depression dataset (Original data)

ABSTRACT

This paper presents Modern Standard Arabic data for the automatic estimation of the risk of depression for online personas based on their daily Arabic tweets. The data were collected from 1-1-2020 to 1-1-2021 using automatically collected samples of depression and non-depression tweets. The data contain 1229 records. These data can be used to develop machine-learning tools to identify the risk of an individual being depressed and to build recommender systems that monitor depression.

© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

* Corresponding author.

E-mail address: aomaghragy@uqu.edu.sa (A. Maghraby).
Social media: ⬅️@aomaghragy_m (A. Maghraby), ⬅️@ii_hes1 (H. Ali)

https://doi.org/10.1016/j.dib.2022.107999
2352-3409/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)
### Specifications Table

| Subject                       | Data science |
|-------------------------------|--------------|
| Specific-subject area         | Data science |
| Type of data                  | .csv file    |
| How data were acquired        | Using TWINT tool from python |
| Data format                   | Raw data (.csv) |
| Description of data collection| The raw data of Twitter posts were downloaded using the TWINT tool from Python. The collection process was conducted from 1-1-2020 to 1-1-2021. An Arabic version of the Patient Health Questionnaire (PHQ9) scale was used to specify the categories of the depression dataset. The categories are: (1) losing interest or pleasure in activities, (2) low mood, (3) sleep disorder, (4) loss of energy, (5) weight disorder, (6) feelings of worthlessness, (7) diminished ability to think or concentrate, (8) psychomotor agitation or retardation, and (9) suicidality. Keywords (terms or phrases) that might fall under each of the categories were created, including terms or phrases a person would use to describe their moods, anxiety, and sleep modules, such as "درايتي" and "لايرني", "درايتي" and "لايرني". These keywords were then used to collect relevant tweets from Twitter. |
| Data source location          | Middle East  |
| Data accessibility            | Repository name: Mendeley data |
|                              | Data identification number: 10.17632/myrb2gky8w.1 |
|                              | Direct URL to the dataset: https://data.mendeley.com/datasets/myrb2gky8w/draft?a=660d1700-1694-4475-8a9a-ed50f538b99d |

### Value of the Data

- This dataset contains information about real users of Arabic tweets to acquire knowledge about a person’s psychology. It can be used for the early detection of depression types for rapid intervention, which can potentially reduce the suffering from the disorder and the stigma associated with mental illness.
- This dataset provides a unique opportunity for researchers, primary care clinics, and psychologists to detect users’ depression early (12 different mental health disorders, including depressed mood, loss of interest, change in appetite or weight, feeling guilty or worthless, sleep disorder, psychomotor agitation or impairment, fatigue or loss of energy, poor concentration and persistent thoughts of death or suicidal ideation) based on their last 14 days of tweets.
- This dataset can be used to automatically complete the Patient Health Questionnaire (PHQ-9), the most well-known psychometric symptoms report for measuring depression symptoms and mental illnesses [1].
- This dataset can help researchers easily build machine-learning and classification models to detect users’ depression types.
- This is a language-dependent dataset designed for the Arabic language with language-dependent features. However, researchers can recreate a similar dataset for other languages, such as English, by following the same steps explained in this work.

### 1. Data Description

This paper presents Modern Standard Arabic data for the automatic estimation of the risk of depression for online personas based on their daily Arabic tweets. The data were collected from 1-1-2020 to 1-1-2021 using automatically collected samples of depression and non-depression tweets. The final data contain 1229 records. These data can be used to develop machine-learning tools to identify the risk of an individual being depressed and to build recommender systems.
that monitor depression. Fig. 1 shows the final dataset overview, and Fig. 2 shows a sample from the final dataset. Fig. 3 shows an English version of Fig. 2.

### 2. Experimental Design and Materials and Methods

An Arabic version of the Patient Health Questionnaire (PHQ-9) scale [1], a well-validated measurement tool for depression, was used in this manuscript. Although it cannot be used formally as a standalone assessment to diagnose clinical depression, it can be used as the first step to screen and detect depression [2,3]. It is commonly used in depressive symptoms self-administered identification tools, online design studies, research, and psychophysical clinics during the first appointment of subjects.

Subjects are required by PHQ-9 to respond by one of the following numbers: 0, 1, 2, or 3 to nine psychophysical questions that indicate depression. Each of these numbers indicates a certain degree of depression, as shown in Table 2. The PHQ9 scale questions are shown in Table 1. In these questions, users are asked about their psychophysical state in the last two weeks (14 days), emotions (e.g., feeling down, tired, tired, anxious), appetite, the occurrence of any sleeping disorder, slowness in movement, and suicidal thoughts [1].

PHQ-9 scores have been shown in previous studies to be strongly correlated with the results of psychologist depression diagnosis in the psychophysical clinic [4]. Generally, subjects with scores ranging from 0 to 9 are considered healthy, subjects with scores ranging from 10 to 19 are considered to have mild depression, and subjects with scores of 20 points or above are considered severely depressed.

Although depression detection is the first step in treating it [2], more than a quarter of most depression patients remain undiagnosed [3]. The main issue is that most patients dislike traditional depression detection methods, such as the paper version of the PHQ-9 scale. In summary, subjects find the questionnaire intrusive, cumbersome, and feared [5].

Twitter data were collected in this work to deal with this issue. It used to automatically answer the nine questions of PHQ-9 for depression screening and detection. Researchers can
| tweet in English                                      | category                                      |
|-----------------------------------------------------|-----------------------------------------------|
| Peace to those who sit alone, scattered with thoughts scattered with cant think | Diminished ability to think or concentrate     |
| I am unable to focus a lot of distraction           | Diminished ability to think or concentrate     |
| I would like to scream, I am disappointed           | Feelings of worthlessness                     |
| I let myself down                                   | Feelings of worthlessness                     |
| I get totally pissed off and i would like to speak with anyone | losing interest or pleasure in activities    |
| You are right, and loneliness is boredom            | losing interest or pleasure in activities     |
| very, very much and forever tired                   | loss of energy                                |
| Why do I feel like everyone is ignoring me          | loss of energy                                |
| I feel tight                                        | low mood                                      |
| I am sad and my weight is gaining                   | low mood                                      |
| I feel a strong lethargy                            | Psychomotor agitation or retardation           |
| I want to sleep, wake up, get excited, come back, enough for a week of sleep | sleep disorder                                |
| my sleeping pattern is messed up                     | Psychomotor agitation or retardation           |
| I swear to God it’s been a long time I didn’t sleep well I miss sleeping | sleep disorder                                |
| I really want to die now                            | Suicidality                                   |
| I want to die but without pain                      | Suicidality                                   |
| To be honest, I don’t have a desire to eat          | weight disorder                               |
| The last meal I ate was lunch At that time I was not able to eat anything | weight disorder                               |

Fig. 2. Sample of the final dataset.

| tweet in Arabic                                      | category                                      |
|-----------------------------------------------------|-----------------------------------------------|
| Diminished ability to think or concentrate           | Diminished ability to think or concentrate     |
| Feelings of worthlessness                            | Feelings of worthlessness                     |
| losing interest or pleasure in activities            | losing interest or pleasure in activities     |
| low mood                                            | low mood                                      |
| Psychomotor agitation or retardation                 | Psychomotor agitation or retardation           |
| Sleep disorder                                      | Sleep disorder                                |
| Suicidality                                         | Suicidality                                   |
| weight disorder                                     | weight disorder                               |

Fig. 3. Sample of the final dataset in English.
Table 1
Patient health questionnaire-9 (PHQ-9).

| No. | English/Arabic Questions                                      | 0 | 1 | 2 | 3 |
|-----|--------------------------------------------------------------|---|---|---|---|
| 1   | Little interest or pleasure in doing things?                |   |   |   |   |
| 2   | Feeling down, depressed, or hopeless?                       |   |   |   |   |
| 3   | Trouble falling or staying asleep, or sleeping too much?    |   |   |   |   |
| 4   | Feeling tired or having little energy?                      |   |   |   |   |
| 5   | Poor appetite or overeating?                                |   |   |   |   |
| 6   | Feeling bad about yourself — or that you are a failure or have let yourself or your family down? |   |   |   |   |
| 7   | Trouble concentrating on things, such as reading the newspaper or watching television? |   |   |   |   |
| 8   | Moving or speaking so slowly that other people could have noticed? Or so fidgety or restless that you have been moving a lot more than usual? |   |   |   |   |
| 9   | Thoughts that you would be better off dead, or thoughts of hurting yourself in some way? |   |   |   |   |

Table 2
Valid values for PHQ-9.

| Value | Value Meaning | Description | Display Order |
|-------|---------------|-------------|---------------|
| 0     | Not at all    | A subjective answer of non-agreement. | 0 |
| 1     | Several days  | An indefinite small quantity that is more than two but less than many. | 1 |
| 2     | More than 1/2 the days | A time period that includes one-half of the total number of days plus at least one more. | 2 |
| 3     | Nearly every day | Not quite or just short of: Including all possibilities or occurring at regular intervals. | 3 |

use this dataset to develop an automatic mental health assessment methodology in which the depression levels of the subjects are detected.

This work began by creating categories for each of these questions, as shown in Table 3. We then chose keywords (terms or phrases) that could fall under each of the categories. The keywords included terms or phrases a person would use to describe their moods, anxiety, and sleep modules, as shown in Table 4. The individual to be diagnosed with depression must experience five or more symptoms during the last two-week (14 days) period of the following nine categories: losing interest or pleasure in activities, low mood, sleep disorder, weight disorder, loss of energy, feelings of worthlessness, diminished ability to think or concentrate, psychomotor agitation or retardation, and suicidality [1]. These keywords were used to collect relevant tweets from Twitter using the TWINT library as follows [6]:

- Python programming language with the TWINT library was used to collect relevant tweets.
- Up to 10000 tweets were scraped for each category.
- Tweets were generated by 38101 Saudi Twitter users
- Data were scraped from 1-1-2020 to 1-1-2021.
- Records with multiple attributes (id, conversation_id, created_at, date, time, timezone, user_id, username, name, place, tweet, language, mentions, urls, photos, replies_count, retweets_count, likes_count, hashtags, cashtags, link, retweet, quote_url, video, thumbnail,
Table 3
PHQ-9 categories.

| No. | English PHQ9 Categories                           | Arabic PHQ9 Categories                          |
|-----|---------------------------------------------------|------------------------------------------------|
| 1   | Losing interest or pleasure in activities         | لعم داب يب اوبقٍ لاب شريةب فاحثباتلإا، ثلق وابعبٍ ثلق |
| 2   | Low mood                                          | مابرابٍ واب رعبٍ ثقيق واب رحبٍ روجٍ رفغٍ رفغٍ |
| 3   | Sleep disorder                                    | دادجٍ رٍ رشلا بوربا واب قلطم ومن واب بوربا وفُرُفُرٍ |
| 4   | Loss of energy                                    | قوّطلا رٍ دادجٍ رفغٍ رفغٍ لفسيحٍ واب رحبٍ روجٍ رفغٍ |
| 5   | Weight disorder                                   | دامبلأا رٍ بٍ الباطٍ لبونٍ وفُرٍ وفُرٍ وفُرٍ |
| 6   | Feelings of worthlessness                         | انتقحاوٍ واب لفسيحٍ بثداشٍ ثلق لفسيحٍ روجٍ رفغٍ رفغٍ |
| 7   | Diminished ability to think or concentrate        | نويرفغٍ رفغٍ رفغٍ واب نويرفغٍ رفغٍ رفغٍ بثداشٍ ثلق لفسيحٍ |
| 8   | Psychomotor agitation or retardation              | واب نيرفغٍ رفغٍ رفغٍ واب نويرفغٍ رفغٍ رفغٍ وفُرٍ وفُرٍ |
| 9   | Suicidality                                       | نويرفغٍ رفغٍ رفغٍ واب نويرفغٍ رفغٍ رفغٍ وفُرٍ وفُرٍ |

near, geo, source, user_rt_id, user_rt, retweet_id, reply_to, retweet_date, translate, trans_src, trans_dest were scrapped.

The final dataset includes 48,821 records Table 5. shows the number of collected recorders for each category.

1.1. Pre-processing dataset

In this stage, three steps were followed: (1) labeling dataset, (2) cleaning dataset, and (3) reviewing dataset.

1.1.1. Labeling the dataset

Each record is labeled by one category name (there are nine categories, as shown in Table 3).

1.1.2. Cleaning the dataset

Cleaning is an essential step in almost any Natural Language Processing (NLP) task. It aims to eliminate incomplete, noisy, and inconsistent data. The dataset was analyzed before cleaning using Pandas Profile Report tools Fig. 4. shows the results of this analysis. According to this figure, the biggest problem with this dataset is missing cells (missing values) at 34.9%. Therefore, two cleaning steps were followed.

Cleaning the dataset file:

- **Removing empty columns**: removed 10 empty columns that did not contain any value.
- **Removing unwanted columns**: removed 24 unwanted columns that did not contain any important value.
- **Removing duplicated records**: removed duplicated rows (0.04%).

Fig. 5 shows the dataset overview after cleaning the dataset file.
I cannot focus - My thinking is scattered - My attention is distracted

I feel lethargic - I feel lazy - I am lazy - I have laziness - I have a sluggish tongue

I want to die - I want to commit suicide - I do not want to live - I want to end my life - I want myself to die - Oh Lord I die - Oh Lord my end my life - I wish I had died - I wish I was dead

Table 4
Keywords for the PHQ-9 categories.

| Categories No. | Keywords |
|---------------|----------|
| 1             | Talim - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم - لم يتم |
| 2             | قهرا - توما - توما - توما - توما - توما - توما - توما |
| 3             | يف - في البداية - في البداية - في البداية - في البداية - في البداية - في البداية - في البداية - في البداية - في البداية - في البداية |
| 4             | I want - I want - I want - I want - I want - I want - I want - I want - I want - I want - I want |
| 5             | I am - I am - I am - I am - I am - I am - I am - I am - I am - I am - I am |
| 6             | مبص - مبص - مبص - مبص - مبص - مبص - مبص - مبص |
| 7             | الى - الى - الى - الى - الى - الى - الى - الى - الى - الى - الى |
| 8             | I feel - I feel - I feel - I feel - I feel - I feel - I feel - I feel - I feel - I feel - I feel |
| 9             | I want - I want - I want - I want - I want - I want - I want - I want - I want - I want - I want |

Table 5
No. of tweets under each category.

| No. | Categories                               | No. of records |
|-----|------------------------------------------|----------------|
| 1   | Losing interest or pleasure in activities | 10012          |
| 2   | Low mood                                 | 8019           |
| 3   | Sleep disorder                           | 7347           |
| 4   | Loss of energy                           | 9705           |
| 5   | Weight disorder                          | 319            |
| 6   | Feelings of worthlessness                | 2068           |
| 7   | Diminished ability to think or concentrate | 3120         |
| 8   | Psychomotor agitation or retardation      | 6162           |
| 9   | Suicidality                              | 2069           |

Total records 48,821 records
### Dataset statistics

|                         |                         |
|-------------------------|-------------------------|
| Number of variables     | 37                      |
| Number of observations  | 48826                   |
| Missing cells           | 629774                  |
| Missing cells (%)       | 34.9%                   |
| Duplicate rows          | 0                       |
| Duplicate rows (%)      | 0.0%                    |
| Total size in memory    | 13.8 MiB                |
| Average record size in memory | 296.0 B               |

### Variable types

|                  |              |
|------------------|--------------|
| Unsupported      | 13           |
| Categorical      | 20           |
| Numeric          | 4            |

**Fig. 4.** Overview of the dataset before cleaning.

### Dataset statistics

|                         |                         |
|-------------------------|-------------------------|
| Number of variables     | 4                        |
| Number of observations  | 44362                   |
| Missing cells           | 3                        |
| Missing cells (%)       | < 0.1%                   |
| Duplicate rows          | 0                        |
| Duplicate rows (%)      | 0.0%                     |
| Total size in memory    | 1.4 MiB                  |
| Average record size in memory | 32.0 B              |

### Variable types

|                  |              |
|------------------|--------------|
| Numeric          | 2            |
| Categorical      | 2            |

**Fig. 5.** Overview of the dataset after cleaning.

### Cleaning tweets:

- **Removing URLs:** removed Tweets links that did not contribute to a depression classification.
- **Removing usernames:** Usernames (@user) were also removed from the tweets.
- **Removing duplicated letters:** replaced any letter that appeared consecutively more than two times in a word with one letter.
- **Removing punctuation:** removed punctuation, such as full stop, comma, and brackets.
- **Removing stop words:** removed the most common words in the Arabic language (articles, prepositions, pronouns, conjunctions, etc.) that did not add much information to the text.
- **Removing emojis:** removed all emojis from tweets.
- **Removing English words:** removed all English words.
- **Text normalization:** Transformed a text to a unified form; removed Al-tashkil and elongation.

### 1.1.3. Reviewing the dataset

Following the cleaning step, each record in the dataset was manually reviewed. Any records that did not relate to depression (e.g., records not describing personal moods, anxiety, movement, sleep mode, suicidal thoughts, or records not associated with the PHQ-9 questions, as shown in Table 1) were disregarded. The number of examples for each category/class label...
was also balanced to achieve the right accuracy when using machine-learning classification algorithms. The final dataset contains 1229 records of 928 Saudi Twitter users as showing in Fig. 1.

Ethics Statement

Our data does not provide any personally identifiable information and only the tweet IDs and human annotated stance labels are shared. Thus, all data are fully anonymized and were collected and distributed under Twitter’s Developer Policy 2021 [7].

The PHQ-9 is available to healthcare providers completely free of charge and the legal copyright holder, explicitly states that “no permission [is] required to reproduce, translate, display or distribute [the PHQ-9]” [8].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Data Availability

Modern Standard Arabic mood changing and depression dataset (Original data) (Mendeley Data).

CRediT Author Statement

Ashwag Maghraby: Conceptualization, Methodology, Validation, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Supervision; Hosnia Ali: Data curation, Investigation, Resources, Writing – original draft.

Acknowledgment

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] National Institute on Drug Abusel, instrument: Patient Health Questionnaire-9 (PHQ-9), National Institute on Drug Abuse, 2019 https://cde.drugabuse.gov/instrument/f226b1a0-897c-de2a-e040-bb89ad433899. Accessed December 1, 2019.
[2] A.L. Siu, K. Bibbins-Domingo, D.C. Grossman, L.C. Baumann, K.W. Davidson, M. Ebell, et al., Screening for depression in adults: us preventive services task force recommendation statement, Jama 315 (2016) 380–387.
[3] R.M. Epstein, P.R. Duberstein, M.D. Feldman, A.B. Rochlen, R.A. Bell, R.L. Kravitz, et al., I didn’t know what was wrong: how people with undiagnosed depression recognize, name and explain their distress, J. Gen. Intern. Med. 25 (2010) 954–961.
[4] K. Kroenke, R.L. Spitzer, J.B. Williams, The phq–9: validity of a brief depression severity measure, J. Gen. Intern. Med. 16 (2001) 606–613.
[5] M.D. Weist, M. Rubin, E. Moore, S. Adelsheim, G. Wrobel, Mental health screening in schools, J. Sch. Health 77 (2007) 53–58.
[6] GitHub, (2020), twintproject/twint, https://github.com/twintproject/twint. Accessed January 1, 2020.
[7] Twitter, Developer agreement and policy, (2021). https://developer.twitter.com/en/developer-terms/agreement-and-policy. Accessed January 1, 2021.
[8] BMED report, health and wellness through psychological science, (2021), Patient Health Questionnaire –9 is an excellent free psychological screening instrument for depression, Lightning Base. https://www.bmedreport.com/archives/14638. Accessed January 1, 2021.