| | |
|---|---|
| Manuscript Submitted | 12.10.2022 |
| Accepted | 30.12.2022 |
| Published | 31.12.2022 |

# The Framework of Data Preparation for Mental Health Detection on Twitter

**Rohizah Abd Rahman**
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
*rohizah@ukm.edu.my*

**Mohd Shahrul Nizam Mohd Danuri**
Faculty of Science & Information Technology
Kolej Universiti Islam Antarabangsa Selangor
*msnizam@kuis.edu.my*

**Azzan Amin**
The Lorry Online Sdn. Bhd. Malaysia
*azzan@thelorry.com*

**Abstract**

*This study aims to generate a framework for data preparation for mental health detection on Twitter. The process of data preparation for mental health detection needs a particular stride. Twitter provides massive data with rich content from Online Social Networks (OSNs). However, this data consists of various types: text, audio, image, and video. Moreover, Twitter's tweets involve hashtags, URLs, retweets, and mentions, but the tweets are meaningful. The tweets also consist of noisy, inconsistent data, anomalies, incomplete data, short-form words, and unmeaningful data. A data preparation framework for mental health detection was proposed to solve these problems. The framework involved Data Collection and Extraction steps, Expert Manual Annotation, Text Cleaning, and Text Representation. Each framework's process was conducted through the experimental method using the Python language. The data was collected for a total of 19,744 tweets in English related to mental health problems for one week. Hence, this study is related to mental health problems, and the manual annotation needs an expert. The expert will annotate the clean text data to detect mental health problems. The text representation was conducted using N-Grams, TFIDF, Bag of Words, and Lemma. These methods represent data for modelling using machine learning techniques. Using the framework could become a process for other problem detection for data preparation. Data preparation is essential for efficient data modelling.*

***Keywords*:** *Data Preparation, Mental Health Detection, Twitter.*

## 1. Introduction

Globally, an estimated 450 million people worldwide suffer from mental health problems, accounting for 13 percent of the world's disease burden (Global Burden of Disease Study 2013 Collaborators 2015). The World Health Organisation estimates that 1 in 4 individuals will experience a mental disorder at some point in their lives (World Health Organization 2001). Recently, studies on mental health detection for Online Social Network (OSN) users have been conducted by researchers (Gedam & Paul, 2021; Kim et al., 2021; Lin et al., 2017; O'Dea et al., 2015; Tsugawa et al., 2015). The data on mental health problems was collected through OSNs, such as Facebook, Twitter, Weibo, and YouTube. The data from OSNs is massive and meaningful.

Aside from mental health detections, data also is imperative for specific purposes, such as sales data, weather, forecasting, sentiment, and politics. Massively large amounts of data are accessed through many data sources, such as organisations and the web. Besides, the data from Online Social Networks (OSNs), including Facebook, Twitter, WhatsApp, and YouTube, provides massive data, which is valuable to analyse for meaningful information. For example, Twitter's users generate over 400 million tweets daily (Kumar et al. 2013).

Primarily, data from OSNs is in an un-normal form, such as noisy, inconsistent data, anomalies, incomplete data, and unmeaningful data (Zhang et al. 2003). Sometimes, data in OSNs is written in short-form words or mixed languages. This form of data is hard to analyse for meaningful data. The raw data from the OSNs needs to be cleaned to ensure the final data for modelling is meaningful.

Therefore, the data preparation framework for mental health detection on Twitter is recommended to overcome the problems. The framework details are explained under the methodology sub-topic.

## 2. Current Works

### 2.1 Online Social Network

One of the data sources that can access data quickly is the internet. Social media data is formed with various modalities, for example, text, image, voice, and social interaction (Guo et al. 2016). The various OSNs, such as Facebook, YouTube, Twitter, Instagram, and WhatsApp, have become the primary medium of daily communication and are the easiest way to share information (Chang et al. 2016). In addition, OSNs are new communication mediums that have offered Web 2.0-based communication methods to encourage active users (Al-Garadi et al., 2016; Nasution & Noah, 2010; Othman & Danuri, 2016).

### 2.2 Data and Twitter

Data preparation is a crucial phase before analysis. The low-quality data is available from various sources in the OSNs. However, the data form is incomplete, noisy, and inconsistent (Zhang et al., 2003). Users on Twitter generate over 400 million tweets daily, and Twitter users' tweets are also known as "status messages". A tweet can be at most 140 characters in length (Kumar et al., 2013). Twitter's tweets involve hashtags, URLs, retweets, and mentions.

### 2.3 Mental Health Detection

The World Health Organisation (WHO) has defined mental health as a state of well-being in which individuals realise their abilities, can cope with the everyday stresses of life, work productively, and contribute to society (World Health Organization, 2004). However, today's changing social landscape has increased mental health problems and psychological disorders. Individual lifestyle factors, including work pressure, a poor financial situation, family problems, domestic violence, and environmental factors, are likely to have an impact on mental health (Rahman et al., 2018; Robert-McComb et al., 2015). Recently, research on mental health detection through OSNs has been conducted.

Researchers conducted studies related to mental health problem detection, involving stress (Lin et al., 2017), depression (Tsugawa et al., 2015), and suicide (O'Dea et al. 2015).

## 3. Methodology

The development of this study will be implemented using experimental methods. A series of experiments were conducted using Python. The processes involved are Data Collection and Extraction, Expert Manual Annotation, Text Cleaning, and Text Representation.

### 3.1 The Framework of Data Preparation for Mental Health Detection

Figure 1 is the framework for data preparation for mental health detection. The framework involved Data Collection and Extraction, Expert Manual Annotation, Text Cleaning, and Text Representation. The data collection process was conducted using Python as an IDE to harvest data from Twitter for a specific duration. The data preparation details in this framework are explained in the sub-topic finding and discussion.

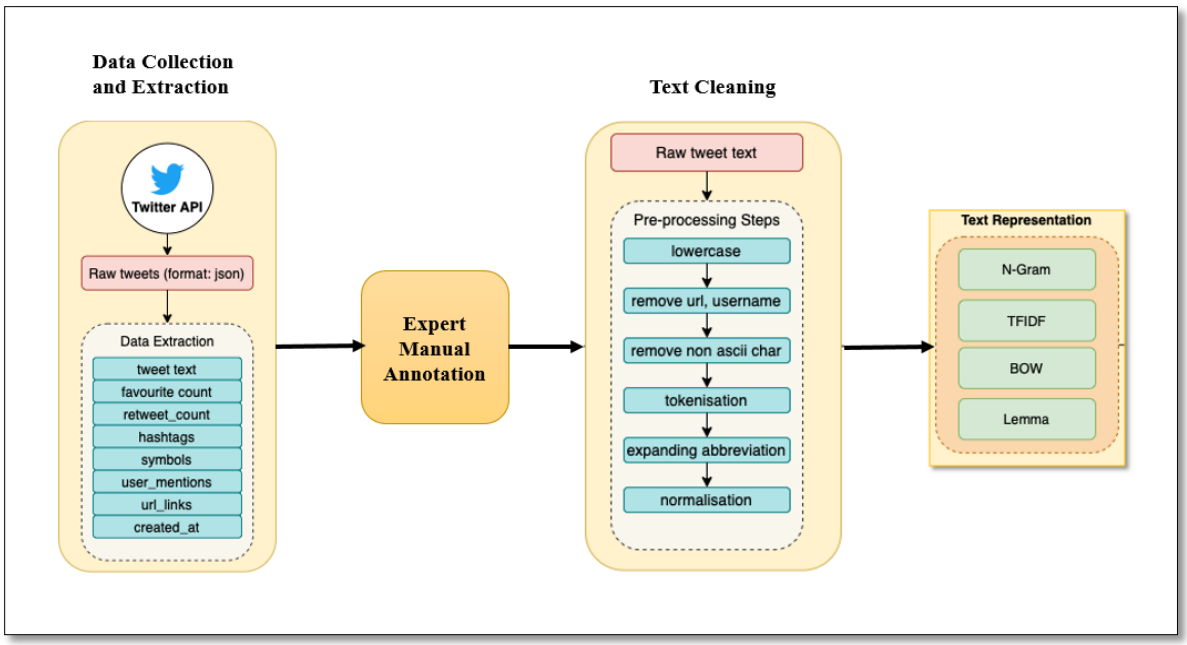


Figure 1: The Framework of Data Preparation for Mental Health Detection

### 3.2 Twitter's Data Collection

The data was collected for a total number of 19,744 tweets in English for one week. The data collected from the Twitter API application requires permission and application from Twitter even though the data is accessible publicly.

### 3.3 Keywords and Constraints

Recently, other researchers have also used keywords in their studies to collect data (Deshpande & Rao, 2017; O'Dea et al., 2015). Keywords are vital in researching mental health problem detection. Researchers should pursue expert advice to ensure those appropriate keywords (Rahman et al., 2020). Figure 2 shows the sample keywords used in this study and asks for an expert. The keywords related to mental health are *'stress,' 'depression,' mental disorder,' delusions, 'sad,' 'anxiety,'* and *'insane.'* This study has some constraints whereby the data collected is suitable for English only and the location focuses on Malaysia.

```
In [135]: # Mental Health Keywords
          keyword_list = ['stress', 'depression', 'insanity', 'mental disorder', 'delusions','sad','anxiety','insane']

          # Malaysia Latitude & Longitude
          lat=3.140853
          lng=101.693207
          radius_km=700
```

Figure 2: The Keywords and Constraints in Data Collection

## 4. Finding and Discussion

Several processes involved in this framework have been done in this study, which are the processes of Data Collection and Extraction, Expert Manual Annotation, Text Cleaning, and Text Presentation. The process details involved are explained in sub-topics 4.1 to 4.4.

### 4.1 The Process of Data Collection and Extraction

*a) Data Collection*

A total number of 19,744 tweets were collected using English. The data was collected for one week with specific keywords related to mental health. The keywords are mentioned in detail in sub-topic 3.2. The Twitter API application requires permission and application from Twitter. It is necessary to set the constraints, such as duration, keywords, language, and location, early before harvesting the data from Twitter. The experiment for harvesting data used Python as the IDE and '*tweepy*' and '*pandas*' as the library. Figure 3 shows the raw data collected from Twitter.

```
In [129]: response_list = get_tweets_save_to_list(tweets_obj_1)
          increase their pain.

          Yoga is not some magical cure-all. It doesn't even help short-term pain, like broken bones.
          Got burns? Cancer? Herniated discs? Tree pose!
          11/
          Gua liat liat teume makin stress
          RT                If a child is pushed towards labor in childhood, then from that time he has to face problems like stress.
          The volunteers…
          RT               :ts: i miss having a life without stress.
          RT                If a child is pushed towards labor in childhood, then from that time he has to face problems like stress.
          The volunteers…
                        Don't worry just a minor set back dont stress!, Stressing makes it worse
          RT              xactly, not gonna stress myself bcs of kcharts anymore. Let's enjoy this sudden comeback bestie and keep votin
          g at mama. im…
          RT              Almighty. Grant us relief from stress and worry, knowing You're in control. Take away all the anxiety and gran
          t us peace. He…
          Less stress keep them lines off your face
          fnf community
          ah yeah, that community gave me multiple stress and anxiety over what happened in the (almost) whole entire year
          i genuinely don't hate it but it's just really terrible
```

Figure 3: The Raw Data from Twitter

*b) Data Extraction*

The data extraction in this study is to focus only on text data. The text data will refer to the tweet's text. The tweet's text is meaningful data and needs cleaning and modelling. Figure 4 shows data after the extraction of some features from Twitter. The list below is the features identified that can be extracted:

- tweet's text
- favourite counts
- number of retweets
- date created
- hashtags
- user mentions
- symbols
- URLs
- media (images)

Out[193]:

| | tweet_text | favourite_count | retweet_count | created_at | hashtags | symbols | user_mentions | urls | is_retweet | tweet_length |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | sTRESS | 0 | 0 | Wed Oct 13 17:58:56 +0000 2021 | [] | [] | [] | [] | False | 1 |
| 1 | I feel so stress if i dont taking care them bu... | 0 | 0 | Wed Oct 13 17:47:40 +0000 2021 | [] | [] | [] | [] | False | 18 |
| 2 | night stress | 0 | 0 | Wed Oct 13 17:40:32 +0000 2021 | [] | [] | [] | [] | False | 2 |
| 3 | Idk why must we stress so much in life😭👌 why ... | 0 | 0 | Wed Oct 13 17:30:15 +0000 2021 | [] | [] | [] | [] | False | 31 |
| 4 | KKR or DC, too stress watch this game | 2 | 2 | Wed Oct 13 17:28:04 +0000 2021 | [] | [] | [] | [] | False | 8 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 19739 | @delusions_d am i tripping or did you say "jer... | 1 | 0 | Mon Oct 18 13:27:33 +0000 2021 | [] | [] | ['delusions_d'] | [] | False | 11 |
| 19740 | MY MUSIC TITLES ARE EVIL BE LIKE :\n\nSeal - S... | 0 | 0 | Mon Oct 18 09:56:37 +0000 2021 | [] | [] | [] | [] | False | 55 |
| 19741 | @delusions_d @demagoguerz record | 0 | 0 | Mon Oct 18 07:06:48 +0000 2021 | [] | [] | ['delusions_d', 'demagoguerz'] | [] | False | 3 |
| 19742 | @kalossharon It's all delusions, I'm telling y... | 2 | 0 | Mon Oct 18 03:47:33 +0000 2021 | [] | [] | ['kalossharon'] | [] | False | 38 |

Figure 4: The Data after Extraction of Some Features on Twitter

### 4.2 Expert Manual Annotation

Manual annotation by experts needs specific data, for example, mental health or other data types. The data was annotated by experts on mental health problems and focused on three categories: stress, anxiety, and depression. This study assigned three experts to annotate the data manually from the four scales, which are 0 (Normal), 1 (Mild), 2 (Moderate), and 3 (Severe). The final results for each category were based on the total average of the annotated data by the three experts. Figure 5 shows some of the cleaned data after being annotated by them.

| | Twitter Text | Stress | Anxiety | Depression | word_count | CleanTweet |
|---|---|---|---|---|---|---|
| 0 | stress | 3 - Severe | 0 - Normal | 0 - Normal | 1 | stress |
| 1 | i feel so stress if i dont taking care them bu... | 3 - Severe | 0 - Normal | 1 - Mild | 18 | feel stress taking care get bullshithahahah go... |
| 2 | night stress | 2 - Moderate | 0 - Normal | 0 - Normal | 2 | night stress |
| 3 | idk why must we stress so much in life why do ... | 3 - Severe | 0 - Normal | 2 - Moderate | 31 | dont know must stress much life care people th... |
| 4 | kkr or dc, too stress watch this game | 2 - Moderate | 0 - Normal | 0 - Normal | 8 | kkr care stress watch game |

Figure 5: The Data After Being Annotated by Expert

### 4.3 The Process of Text Cleaning

Text cleaning or preprocessing data text is essential before modelling. The excellent quality of data is important to ensure that the result of the modelling is high. A few steps for text preprocessing include hiding all the text in lowercase, removing URL and username, removing non-ASCII and characters, tokenizing the words, expanding abbreviations, and normalising the text. The experiment that used Python needs to import the nltk library like 'tokenize.' The text preprocessing includes text cleaning, normalisation, and standardisation. Figure 6 shows the process involved before modelling. Figure 7 describes the frequent word mentions in tweets regularly through Word Cloud.

```
#lowercase the word
text = text.lower()

# remove URLs
text = re.sub('https?://[A-Za-z0-9./]+', '', str(text))
text = re.sub(r"www\.[a-z]?\.?(com)+|[a-z]+\.(com)", '', text)

#remove usernames
text = re.sub('@[^\s]+', '', text)

#remove non ASCII characters
text = re.sub(r'[^\x00-\x7F]+',' ', text)

#remove the # in #hashtag
text = re.sub(r'#([^\s]+)', r'\1', text)

#remove non letter character
text = re.sub("[^a-zA-Z]", " ", text)

#remove word end with semicolon
text = re.sub(r'\w+:\s?','',text)

# Word tokenization
words = tok.tokenize(text)

# Expanding contractions
expanded_words = []
for word in words:
    if word in list(contraction_dict.keys()):
        word = contraction_dict[(word).strip()]
    expanded_words.append(word)
text = ' '.join(expanded_words)
```

Figure 6 The Process of Cleaning The Text

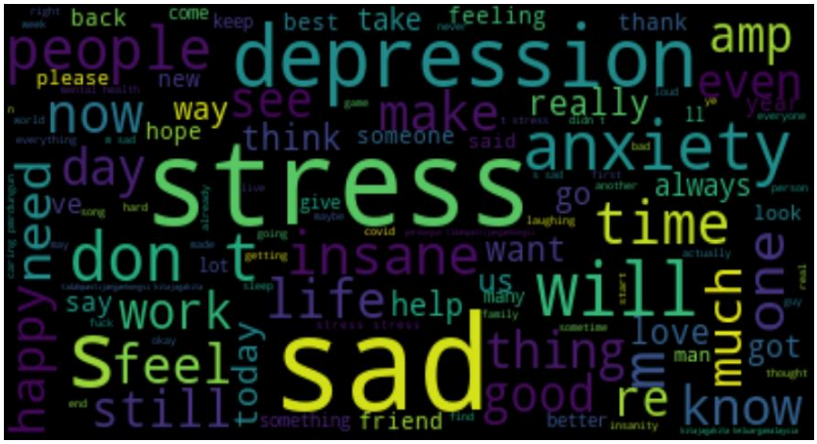


Figure 7 The Word Cloud

### 4.4 The Process of Text Representation

The next step after text preprocessing is text representation. Text representation is a method to convert text from its natural form to vector form. This study used four methods of text representation: N-Gram (Bigram), TFIDF, Bag of Words, and Lemma. The text representation is the way for the data to be ready for modelling using machine learning techniques. Most researchers have selected any of the available machine learning techniques (Ghazali et al., 2013) and a good feature selection will improve the performance of learning algorithms (Al-Moslmi et al., 2015).

*a) Bag of Words*

The Bag of Words (BoW) is a Natural Language Processing (NLP) technique of text modelling and feature extraction technique with text data. This approach is a simple and flexible way of extracting features from documents. The BoW describes the occurrence of words within a document by simplifying the form of text representation into numbers. This process is needed as the algorithm applied in NLP will be done through numbers and not the text itself. Therefore, this research will convert the text with mental health sentiment on Twitter into the BoW.

*b) TFIDF*
The TFIDF is a way of determining a word's relative importance to understand how often the word appears across multiple documents in the field of information retrieval. The words relevant to a specific topic on mental health are likely to occur in documents and much less in other related issues through weighting schemes. On the other hand, less meaningful words will be shared across documents about any subject. Multiple documents need to measure a word's document frequency and numeric statistical analysis.

TFIDF is widely used in analysing search engines, translation, and other text data weighting schemes. This process helps determine relevant words to mental health available in the set of data extracted from Twitter.

*c) N-Grams: Bigram*
N-grams are continuous sequences of words in data. N-grams are used for developing a language model into Unigram, Bigram, and Trigram models. This study used Bigram as a model N-gram. The significance of using this model is that it creates meaningful data. For example, the words *'stress'* and *'not stress*' will be assigned to different class labels. Therefore, the researcher believed using Bigram was enough for this mental health detection research.

*d) Lemma*
Lemmatisation is a text normalisation technique that switches any word to its base root mode. Lemmatisation is responsible for grouping different inflected forms of words into their root form, which has the same meaning. Since the researcher chooses Bigram, the lemmatisation effect on the result is insignificant, especially in mental health detection.

## 5. Conclusion and Future Recommendations
This study offers the framework for data preparation for mental health detection on Twitter. The framework aligns the process involved with Data Collection and Extraction, Expert Manual Annotation, Text Cleaning, and Text Representation. Multiple studies on mental health problems can be done in the future. Research can focus on mental health detection using the native language, not restricted to English only. Examples of issues raised, such as anomalies in collected data, but meaningful and short-form words, also need to be studied in the future. Meanwhile, the N-Gram, BoW, and TFIDF are common processes in text representation. In the future of mental health research, the researcher may use other text representations, such as Word Embedding (Word2Vec, GloVe, or FastText) or Tensor before testing it with any suitable classifier using Machine Learning.

Text cleaning is tedious and critical as this preprocessing will determine the subsequent analysis and results. Adding the expert manual annotation will make the preprocess reliable and validated. Therefore, this study proposed a new data preparation framework for mental health detection in this study.

**Acknowledgement**
Thank you to Universiti Kebangsaan Malaysia for supporting this work under Grant GGPP-2019-004 and Twitter for the data provided.

**References**


Al-Garadi, M. A., Khan, M. S., Varathan, K. D., Mujtaba, G., & Al-Kabsi, A. M. (2016). Using online social networks to track a pandemic: A systematic review. *Journal of Biomedical Informatics*, *62*, 1–11. https://doi.org/10.1016/j.jbi.2016.05.005

Al-Moslmi, T., Gaber, S., Al-Shabi, A., Albared, M., & Omar, N. (2015). Feature selection methods effects on machine learning approaches in Malay sentiment analysis. *1st ICRIL-International Conference on Innovation in Science and Technology (LICIST 2015)*, *October*, 2–5.

Chang, C.-H., Saravia, E., & Chen, Y.-S. (2016). Subconscious Crowdsourcing: A Feasible Data Collection Mechanism for Mental Disorder Detection on Social Media. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 374–379. https://doi.org/10.1109/ASONAM.2016.7752261

Gedam, S., & Paul, S. (2021). A Review on Mental Stress Detection Using Wearable Sensors and Machine Learning Techniques. In *IEEE Access* (Vol. 9, pp. 84045–84066). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3085502

Ghazali, J., Noah, S. A., & Zakaria, L. (2013). Classification of images for automatic textual annotation: A review of techniques. *Journal of Applied Sciences*, *13*(6), 760–767. https://doi.org/10.3923/jas.2013.760.767

Global Burden of Disease Study 2013 Collaborators. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, *386*(9995), 743–800. https://doi.org/10.1016/S0140-6736(15)60692-4.Global

Guo, Q., Jia, J., Shen, G., Zhang, L., Cai, L., & Yi, Z. (2016). Learning robust uniform features for cross-media social data by using cross autoencoders. *Knowledge-Based Systems*, *102*, 64–75. https://doi.org/10.1016/j.knosys.2016.03.028

Kim, J., Lee, D., & Park, E. (2021). Machine learning for mental health in social media: Bibliometric study. *Journal of Medical Internet Research*, *23*(3). https://doi.org/10.2196/24870

Kumar, S., Morstatter, F., & Liu, H. (2013). Twitter Data Analytics. In *Springer*.

Lin, H., Jia, J., Qiu, J., Zhang, Y., Shen, G., Xie, L., Tang, J., Feng, L., & Chua, T.-S. (2017). Detecting Stress Based on Social Interactions in Social Networks. *IEEE Transactions on Knowledge and Data Engineering*, *29*(9), 1820–1833. https://doi.org/10.1109/TKDE.2017.2686382

Nasution, M. K. M., & Noah, S. A. M. (2010). Extracting Social Networks from Web Documents (Extended Abstract). *1 St National Doctoral Seminar in Artificial Intelligence Technology (CAIT'2010)*, 278–281.

O'Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*, 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Othman, M. K., & Danuri, M. S. N. M. (2016). Proposed conceptual framework of Dengue Active Surveillance System (DASS) in Malaysia. *2016 International Conference on Information and Communication Technology (ICICTM)*, 90–96. https://doi.org/10.1109/ICICTM.2016.7890783

Rahman, R. A., Omar, K., Noah, S. A. M., & Danuri, M. S. N. M. (2018). A survey on mental health detection in Online Social Network. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4–2), 1431–1436.

Robert-McComb, J. J., Casey, S., Kim, Y., Hart, M., Norman, R., & Qian, X. (2015). Experimental Models for Research in Stress and Behavior. *Journal of Behavioral and Brain Science*, *05*(07), 295–305. https://doi.org/10.4236/jbbs.2015.57030

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing Depression from Twitter Activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015)*, 3187–3196. https://doi.org/10.1145/2702123.2702280

World Health Organization. (2001). *The World Health Report 2001: Mental health: New Understanding, New Hope*. World Health Organization.

World Health Organization, W. (2004). Promoting mental health: Concepts, Emerging Evidence, Practice. In *Summary Report*. World Health Organization.

Zhang, S., Zhang, C., & Yang, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, *17*(5–6), 375–381. https://doi.org/10.1080/713827180