International Journal of Information Management Data Insights 2 (2022) 100113

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Machine learning-based proactive social-sensor service for mental health monitoring using twitter data

Shailesh Hinduja [a], Mahbuba Afrin [b], Sajib Mistry [b,*], Aneesh Krishna [b]

[a] School of Computer Science, The University of Sydney, Sydney, NSW 2000, Australia
[b] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Australia

ARTICLE INFO

Keywords:
Twitter
Social sensor
Sentiment analysis
Machine learning
Mental health

ABSTRACT

The social media platforms are considered an ecosystem of social sensors where each social media platform user is treated as a social sensor cloud. To overcome the limitations of large-scale mental health surveillance using traditional health administration systems. Although the existing approaches in the literature provide an online detection of mental illness, these are challenging to apply in early detection. Focusing on the Twitter platform, a generic framework is designed in this paper to support proactive mental health monitoring. Detailed data cleaning and pre-processing of the tweets are offered using regular expressions based on observed patterns to ensure accurate results in sentiment analysis. A machine learning mechanism, especially LSTM, is applied for the early detection of at-risk social sensors based on custom event definitions to overcome the limitations of traditional classifiers. The performance of the proposed mechanism has been experimented with, and it outperforms the existing approaches in terms of accurate prediction.

1. Introduction

According to the American Psychiatric Association, *mental illnesses* are serious health conditions where a person's thinking, emotion or behaviour are changed that deteriorates his/her wellbeing (Chen et al., 2022; Ranna Parekh, 2015). Mental illnesses are associated with distress and/or problems functioning in social, work or family activities. According to the world bank, societies worldwide face an enormous disease burden from mental disorders (Marquez & Dutta, 2017). It also mentioned that *depression* affects 350 million persons and is the single largest contributor to years lived with disability globally. It is the third leading cause of disability burden in the world and is prevalent in almost all countries.

It is challenging to address the large-scale mental health surveillance using traditional health administration systems (Batty, Russ, Stamatakis, & Kivimäki, 2017; Kemp, 2017). Only 50% of the people receive treatments in the developed countries of the world, and even the prescribed drugs may not be effective. The situation is worse in developing countries with growing populations, with a severe shortage of psychiatrists and mental health professionals. People suffering from severe depression and psychological distress are likely to commit *suicide* and are at an increased risk of heart attack or stroke (Ji et al., 2020). Mental illness is not restricted to adults; in fact, reports have shown that the number of students starting even from as early as primary school suffering from severe mental illnesses is increasing at great rates (Zhu et al., 2017).

The challenge that governments and health organizations face is that most people are afraid to talk openly about their mental illness due to fear of being excluded in society (Marquez & Dutta, 2017). Hence the number of persons suffering from mental illness is severely underreported, making it even more *difficult* to provide help and assistance. Even with governments increasing their budgets for mental health, there are not enough facilities and trained mental health professionals to serve the growing population suffering from such illnesses. The traditional mental health management and treatment approach requires doctor visits, one-on-one psychological support and medication (which is costly and can be ineffective considering the accuracy). To address these challenges, we are exploring *innovative solutions* using machine learning algorithms on *social media platforms* for early detection of mental illness.

Social media platforms such as Facebook, Instagram, and Twitter are being used daily by millions of worldwide users to communicate and share information (Agarwal, Chauhan, Kar, & Goyal, 2017; Gupta, Kar, Baabdullah, & Al-Khowaiter, 2018; Sharma, Sanghvi, & Churi, 2022). Key social media usage statistics based on the Digital in 2017 Global Overview (Kemp, 2017) indicate that there were 2.80 billion global social media users as of January 2017; and compared to January 2016, there were 482 million more active social media users, which is a 21% increase in social media platform adoption.

* Corresponding Author.
E-mail addresses: shin0210@uni.sydney.edu.au (S. Hinduja), mahbuba.afrin@curtin.edu.au (M. Afrin), sajib.mistry@curtin.edu.au (S. Mistry), a.krishna@curtin.edu.au (A. Krishna).

https://doi.org/10.1016/j.jjimei.2022.100113
Received 13 October 2021; Received in revised form 19 August 2022; Accepted 19 August 2022

Social media platforms can act as a *low cost*, non-intrusive sensor of behavioural changes of a person. The social media platforms can be considered as an ecosystem of social sensors where each user of the social media platform is treated as a "*social sensor cloud*". The messages posted by the user can be treated as "*social-sensor cloud services*" (Sakaki, Okazaki, & Matsuo, 2010). Analogous to medical instruments, e.g., heart rate monitors which record, plot and alert on the heartbeat patterns in a patient; the social media platform can be considered as a *massive instrument* containing all the messages posted by the users. People use these social networks to share their views and opinions on a diverse range of topics such as News, Science and Technology, Sports, Politics, Health, Crises, Emergencies and many more. People also share their personal life moments and experiences and increase their personal health-related information. Each social-sensor service indicates the mental health state of the user at that point in time. *Analyzing* these services can enable governments and health agencies to understand the overall mental state of certain populations by region and can also alert them on the likelihood of an individual suffering from mental illness.

In this paper, we *focus* on the social-sensor cloud services from the Twitter platform. Twitter is one of the popular social media platforms. It is a micro-blogging service, which allows users to type short messages up to 140 characters and post images, and videos (Wikimedia Foundation, 2022). We propose a generic approach for mental illness detection. It consists of several activities such as curating a dataset of users who mentioned that they had an illness, labelling these tweets as *personal, non-personal* and collecting their past six months' tweets. We perform detailed data cleaning and pre-processing of the tweets using regular expressions based on observed patterns to ensure we get accurate results in sentiment analysis. Based on the sentiment analyzer polarity value for each tweet, we mark the tweet as a positive, negative or neutral tweet and use it later in classification tasks using machine learning to predict the likelihood of a person having a mental illness. The major contributions of this paper are claimed as follows:

- A generic framework is designed to support proactive monitoring in social-sensor cloud services based on social media data, especially Twitter data.
- A detailed data cleaning and pre-processing of the tweets is offered using regular expressions based on observed patterns to ensure accurate results in sentiment analysis.
- A machine learning mechanism is developed for the early detection of at-risk social sensors based on custom event definitions to overcome the limitations of traditional classifiers.
- The performance of the proposed mechanism is evaluated in a working prototype based on social media data.

Over the past few years, the use of social media mining for public health monitoring has been a hot and widely discussed topic (Drydakis, 2021; Holzinger, Langs, Denk, Zatloukal, & Müller, 2019). Most of the research has been mainly focused on early disease outbreak detection (Kim, Lee, Park, & Han, 2020; Lee, Agrawal, & Choudhary, 2013). Natural Language Processing (NLP) algorithms have been used to mine Twitter data in real-time to analyze and predict a disease outbreak and to create health maps (Lee et al., 2013). This information is a valuable resource for public health monitoring and surveillance. Although these approaches provide an "online" detection of mental illness, they are challenging to apply in the "early detection". In this paper, we explore the effectiveness of *sentiment analysis* on the user's past messages for the early detection of mental illness. To the best of our knowledge, machine learning on sentiment analysis for the early detection of mental illness is yet to be explored in existing research.

We have taken a different approach by collecting past tweets of users who initially mentioned that they were suffering from a mental illness such as depression. Our framework and prototype mainly focused on detailed and custom data cleansing and performing sentiment analysis on historical tweets to predict the likelihood of a user who has a mental illness. Also, rather than only using traditional statistical and machine learning classifiers explored, we explored both traditional machine learning algorithms such as SVM for binary classification and the latest deep learning algorithms such as LSTMs for performing sequence-based classification to predict the likelihood of a user suffering from mental illness. Applications of such a framework can be applied in public health surveillance, crime monitoring, counter-terrorism strategies, and cyber-bullying prevention. Experimental results prove the efficiency of the proposed framework.

The rest of the paper is organized as follows. The motivation for using Twitter data is discussed in Section 2. The proposed framework is discussed in detail in Section 3. A working prototype is developed in Section 4 to validate the efficiency of the proposed mechanism. The experiment results are discussed in Section 5. The contributions to literature, and implications for practice are discussed in Section 6. Section 7 concludes the paper.

2. Motivation

Twitter is about "What's happening?". People around the world use Twitter to post messages about events happening around them and within them. These days, most people have a smartphone, and even when they are with their family and friends, it is common to see them busy posting on some form of social media. For people suffering from mental illnesses who cannot speak up openly about their illness to a friend or relative, social media platforms such as Twitter provide them with an environment where they can openly talk about their thoughts, feelings and what's going on in their lives. Tapping into these social media platforms for early detection of the likelihood of a user suffering from a mental illness can help in providing timely assistance and support, which can prevent a mentally ill person from causing harm either to themselves or others. The traditional approach is a reactive method, in which the illness is detected only when a person physically visits the doctor and openly discusses the issue; hence early detection is not possible.

We began with using Twitter's advanced search Twitter (2022) on the following terms: "Alzheimer Asperger Bipolar Depression Dementia". Several types of search results were returned based on tweets from users, such as:

1. *I hear voices all the time. I am afraid that I have depression but I'm afraid to get help.*
2. *I've been diagnosed with depression and having anxiety attacks, yesterday I had the worst anxiety attack.*
3. *I have been diagnosed with depression and I've been dealing with it for years. Its so sad to see people lose their...*
4. *I've been diagnosed with bipolar, depression, and anxiety since I was 12 but I'm gonna get a new psych evaluation*

Looking at the past tweets of these users, we noticed that most of their tweets were sad or relating to their mental illness.

This led us to the intuition to collect the past tweets of users who tweeted that they were suffering or had been diagnosed with a mental illness, to analyze their sentiments on each of their past tweets and to try and discover patterns and features that could be useful to predict the likelihood of a person being mentally ill. To the best of our knowledge, we could not find any such dataset specific to mental illnesses; hence we decided to build our own dataset.

3. Proposed methodology

Our target is to detect the likelihood of a person suffering from mental illness "as early as" possible using social sensor cloud services. It is fundamentally different from "real-time" detection of mental illness. Traditional techniques for real-time detection of mental illness usually perform only NLP techniques on the data collected from social services. For example, such techniques collect messages posted by users on social media platforms in real-time and check for content matching medical

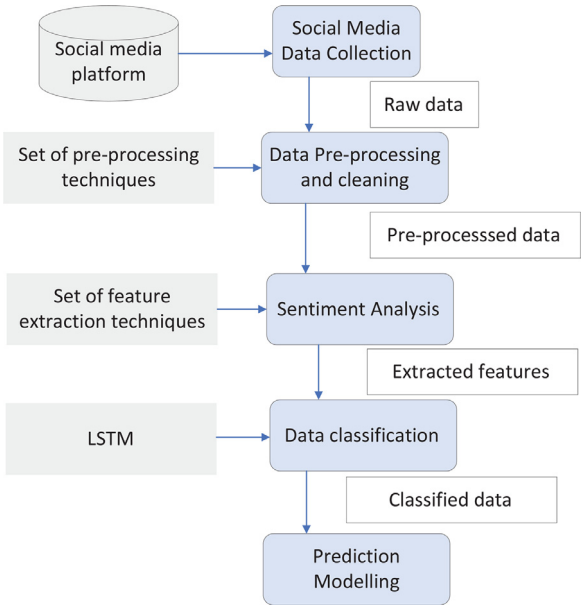

Fig. 1. The mental illness detection framework.

health-related terms in databases to detect users suffering from an illness. This is not applicable in early detection.

Our past behaviours and mindset affect our present health status. Hence, we believe that historical sentiment analysis and machine learning could be a potential approaches for early detection. We propose the following generic framework for implementing an early warning system to detect the likelihood of a person who has a mental illness using social sensor cloud services. Each part of the framework has fundamental research issues. As shown below in Fig. 1, the entire process can be divided into functional blocks, from collecting the data to labelling, data cleaning, performing sentiment analysis, and applying machine learning techniques to classify a user as suffering from mental illness.

3.1. Data collection and pre-processing

The data collected from the social media platform and pre-processing for further analysis depends on the following steps:

3.1.1. Identifying health-related messages and true labelling

Most social media platforms provide APIs to collect messages posted by users. There are over 450 types of mental illnesses ranging from minor to severe, but for our research, we decided to focus on the following: alzheimer's, asperger's, bipolar, dementia, depression, and schizophrenia. To limit the initial set of data to messages related to personal mental illness, we plan to run a query to search for messages posted by users such as: *I suffer from depression, I've been diagnosed with depression, I have been diagnosed with depression, I have depression, been suffering from depression*, and repeat this for the remaining five illnesses listed above. This is to ensure that our initial set of data will be mostly related to users mentioning their personal mental health rather than getting data related to news, advertisements or information posted by health agencies and other charity organizations.

After collecting this initial set of user messages, we classify whether the message was genuinely related to the user's health issue or not. It can be confusing to do this binary classification even for humans because of the nature of the language and usage. For example a user can be sarcastic or joke about an issue, or sometimes it could be a genuine health-related message but not directly relating to the user itself. Health experts are used to classifying each message with a Yes/No (binary classification) - indicating whether the message was truly genuine and related to the user's health. It is considered that if any neutrality or ambivalence situation occurs, it will follow the proposed approach in (Valdivia, Luzón, Cambria, & Herrera, 2018).

Since millions of messages are posted daily on social media platforms; hence our focused approach of collecting these initial sets of messages helps cut out the noise from any other unrelated public messages. By manually labelling these messages, we are further accurately classifying these messages as personal health-related. In future, these manually labelled messages can be used as a training dataset to automatically classify newly downloaded messages using supervised machine learning algorithms as personal health-related or not.

3.1.2. Building historical records

Based on these labelled initial sets of user messages, we plan to collect each user's past messages going back six months or one year, depending on the accessibility and speed of the data collection. To improve the accuracy of supervised machine algorithms in predicting the likelihood of a user to be suffering from a mental health illness, we plan to collect all the messages posted by the initial set of users in this time frame.

The messages posted by users on social media platforms may contain undesired text such as: *urls, username mentions, swearwords with special characters, acronyms*. The messages may even have spelling mistakes. To perform accurate sentiment analysis on these messages, we plan to pre-process them using regular expressions and correct the spellings using existing spelling correction packages.

By collecting all the past user messages and performing data cleansing on them, we will ensure that input to the sentiment analyzers is clean, improving sentiment polarity calculation accuracy.

3.2. Processed data analysis and robustness check

After data cleaning, the pre-processed data are used for sentiment analysis. After the sentiment analysis, data are classified using machine learning technique, specially LSTM.

3.2.1. Polarity generation using sentiment analysis

Sentiment analysis, also known as opinion mining, is basically the process of determining the emotional tone underlying a piece of text. This is done by calculating the polarity of the text based on the number of positive, negative, and neutral words used in the text. The polarity is generally calculated by referring to a lexicon dictionary of words with their polarities, and is a decimal number ranging from −1.0 to 1.0. Polarity values closer to 1.0 indicate positive sentiment, closer to −1.0 indicate negative sentiment, while polarity values equal to 0.0 indicate neutral sentiment.

3.2.2. Selecting appropriate machine learning technique

Based on the sentiment analyzer results on users' past messages, we plan to summarize the results in two types of datasets:

1. User Summary dataset
2. User Weekly Negativity dataset

The User Summary dataset will contain user wise summary of calculated measures such as Average Polarity, Number of Positive tweets, Number of Negative tweets, Number of Neutral tweets, Total number of tweets and any other additional calculations based on these measures. This dataset can then be used for applying traditional supervised machine learning algorithms for binary classification to predict whether a user has a mental illness based on their past sentiments. We plan to use Support vector machines (SVM), a popular and accurate supervised learning algorithm for binary classifying of linearly and non-linearly separable classes of data.

The User Weekly Negativity dataset will contain user-wise weekly negativity values (number of negative tweets by week) and can be used

for sequence-based classification using the latest deep learning techniques. We plan to use LSTM (Long Short-Term Memory), a form of a gated recurrent neural network, to help classify sequences of the number of negative sentiments by week over time by a user as likely to be suffering from mental illness.

The robustness of using machine learning for sentiment analysis and classification is described in detail in the following section.

4. Development of working prototype for evaluation

To evaluate the feasibility of our proposed design framework, we implement a working prototype and experiment the effectiveness of using traditional machine learning algorithms for supervised classification compared with deep learning algorithms using sequence-based classification.

The entire software stack for implementing the framework comprised mainly of: Python, MongoDB (MongoDB, 2018), TwitterAPI (Twitter, 2018), TextBlob (Steven Loria, 2022), Keras (Chollet et al., 2015), Tensorflow (Abadi et al., 2015) and Sci-kit Learn (Pedregosa et al., 2011). We use Python as the main programming language due to its growing popularity in the data science community, integration with the latest deep learning packages and ease of use. We decided to use MongoDB as the database to store all the tweets since it natively stores data in JSON format, and TwitterAPI also returns tweets in JSON format. Also, MongoDB being a NoSQL database, is horizontally scalable and can handle large datasets. For performing sentiment analysis, we use TextBlob for its powerful and easy access to processing large blobs of text and performing various NLP tasks. We decided to use Keras with Tensorflow as the backend engine to explore machine learning and classification using neural networks. The entire experiment was conducted on a Windows 10 desktop with a 3.40 GHz Intel i7-6700 CPU and 16 GB RAM.

4.1. Data corpus creation

4.1.1. Initial data collection

Twitter provides access to user's public tweets via APIs. While it allows access to the full archive going back to 2006, a significant cost is associated with it. Twitter also provides a standard Search API (REST) for researchers to freely access approximately 1% of the daily public tweets for the past seven days. The standard Search API is based on relevance and not on completeness; hence some tweets and users can be missing from the search results.

Using Twitter's standard Search API, we ran a query to search for all tweets that contained the text: *I suffer from depression* or *I've been diagnosed with depression* or *I have been diagnosed with depression* or *I have depression* or *been suffering from depression*, and repeated this for the remaining five illnesses: alzheimer, asperger, bipolar, dementia, and schizophrenia.

This query resulted in 2138 tweets with the breakup as shown in Table 1.

Even for this small data set, we observed that it correlated with the fact that a large proportion of the population suffers from depression.

Table 1
Query breakup of 2138 tweets.

Mental illness	Number of tweets
Depression	1600
Bipolar	191
Alzheimer	135
Dementia	72
Schiznoprenia	50
Asperger	90

Table 2
Data cleaning steps.

Data cleaning criteria	Example	Action taken
Replace matching swearwords	sh*t	'shit
Replace acronyms	gr8	great
Remove more than two repeating characters	helloooo world	hello world
Remove special characters	#'notwell	notwell
Perform spelling correction	dipresion	depression

4.1.2. Data labelling

For these initial 2138 user tweets, we wanted to classify whether the tweet was related to the user's personal health issue or not. Since the list was not very long we decided to ask three persons separately to classify each tweet with a Yes/No (binary classification) - indicating whether the tweet was genuine and related to the user's health. We finally classified each tweet as personal health-related based on a majority vote (Yes/No).

4.1.3. Past tweets data collection

To properly analyse a person's past tweets, we wanted to get the complete set of tweets posted by the user in English for all the days between 01/06/2017 to 30/11/2017. We could not use Twitter's standard Search API as it would give us only a sample of the tweets for the past seven days. Hence, we decided to run a crawler using Scrapy(Kouzis-Loukas, 2016) to collect user's past tweets. We could averagely download approximately 10 to 15 user's tweets based on the above date range. Over the past three months, we downloaded 600 users' past tweets. The tweets were downloaded in raw JSON format to give us the flexibility of importing them into either a relational or a NOSQL database. For our research, the tweets were imported into a MongoDB database instance.

4.1.4. Data cleaning/pre-processing

By randomly observing some of the tweets, we realized that this was going to be an important step before performing sentiment analysis or any further processing. Some of the tweets contained urls, usernames (starting with the "@" character), acronyms (such as "lol","wtf","gr8"), swear words (with and without special characters such as "sh*t"), repeating characters (such as "i looooooovvvvveeeee youu"), emoticons, numbers and other special characters (!@#$%[]", /?). We used regular expressions to detect these text patterns and replace them appropriately. For each tweet, the order in which these text patterns were replaced was also important so as not to lose any important information or emoticons contained in the tweet. Some relevant examples of data cleaning steps are summarized in Table 2.

For the list of acronyms and swear words (Beal, 2016; Sims, 2017), we checked the popular ones used on Twitter and created our own custom list as a reference for data cleaning. Spelling correction was performed using TextBlob (Steven Loria, 2022). The original tweet text was not directly replaced; instead, we created a new column for storing the cleaned tweet text, checking that the text was properly cleaned and leaving room for doing any analysis on the original tweet text.

4.2. Sentiment analysis

There are several sentiment analyzer libraries available, and for our research, we decided to use TextBlob. TextBlob (Steven Loria, 2022) is easy to use and is powerful since it is based on the mighty shoulders of the widely used NLTK and Pattern packages (Lab, 2017; CLiPS, 2017). We applied TextBlob's sentiment analysis on the 979,466 past user tweets to calculate the polarity of the tweets, based on which we then calculated the number of positive, negative and neutral tweets posted by the user, by date. Below are some samples, as shown in Table 3.

While, most of the time, the sentiment analyzer calculated the polarity satisfactorily, the results were subjective at other times. For example, "i am.SO happy" got a polarity score of 0.8, while "WOW IM IN LOVE"

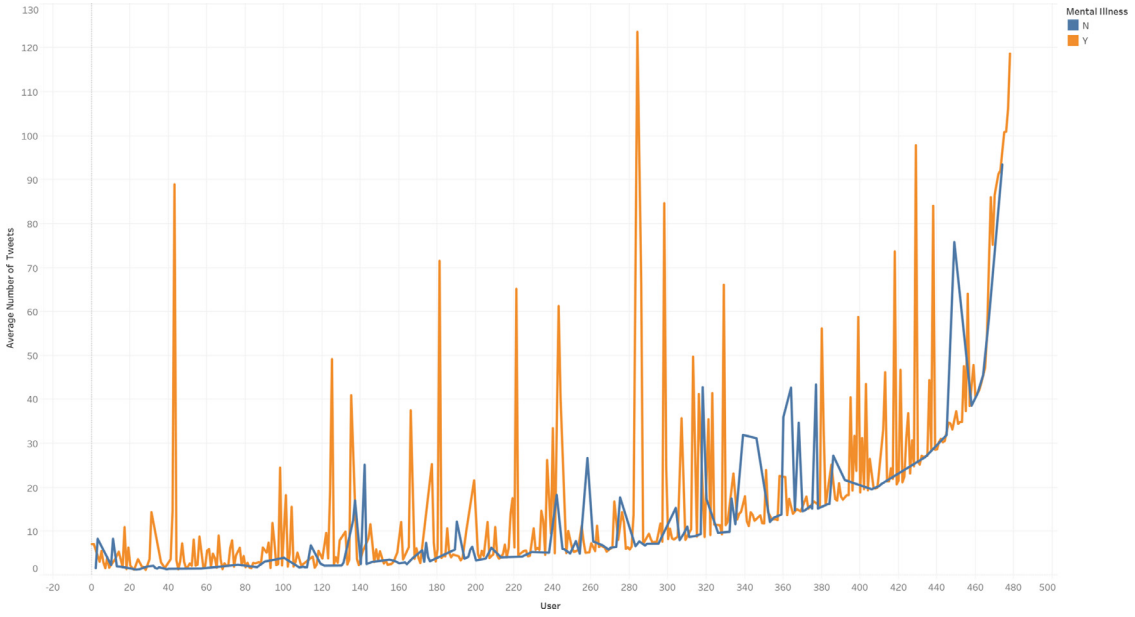

Fig. 2. Average number of tweets per user by class.

Table 3
Sample sentiment polarity.

Cleaned tweet	Polarity
This is very sad. :(Alexa xx	−0.7
How do you hand the dreaded I'm bored	−0.5
You too Evil ?	−1.0
Why am I so tired it's only	−0.2
WOW IM IN LOVE	0.3
i am.SO happy	0.8

got only a score of 0.3, which would have probably been 1.0 if a human being were to give it a score.

4.3. Machine learning classification

4.3.1. Dataset preparation

For the purposes of applying machine learning algorithms, two types of datasets were created based on the sentiment analyzer results: *User Summary* and *User Weekly Negativity*. The User Summary dataset was created to be used for supervised binary classification using traditional machine learning classification algorithms. This dataset consisted of calculated measures based on the polarity of each user's daily activity such as: Average Polarity, Number of Positive tweets, Number of Negative tweets, Number of Neutral tweets, and Total number of tweets. Fig. 2 visualises the dataset distribution. A new measure called the - Negativity Ratio (NR) was also calculated as follows:

$$NR = \frac{Negative\ Tweets}{Positive\ Tweets + Neutral\ Tweets} \quad (1)$$

We observed that users who were labelled as having a personal mental health illness (class: Yes), tweeted more and had a higher number of negative tweets than those labelled as not having a personal mental health illness (class: No). This was encouraging as it indicated that our intuition was right and that the manual labelling of the initial tweets was good.

The User Weekly Negativity dataset was created based on the same sentiment analyzer results for the purpose of doing sequence or time-series-based analysis. Since the Number of Negative tweets was a good indicator for a user labelled as having a mental health illness, we created this dataset by pivoting the number of negative tweets per week for each user. This dataset consisted of user-id, class label and 27-week columns. Fig. 3 visualises the Negativity dataset distribution.

There were some users who had very few tweets ranging from 1 to 49 for the entire six-month period. These users may have either not tweeted much or it is possible that they deleted their tweets. Hence, we did not include them in creating the models. There was also one user who had over 80,000 tweets for the same time period. This could have been a "bot", as it would not be humanly possible to type an average of 436 tweets per day. This was treated as an outlier and hence we did not include this user's tweets in our analysis. Both these datasets finally consisted of 479 users for which we had collected past six months tweets and who had posted at least 50 tweets for this time period. Out of these 479 users, 356 users were labelled as: Yes (indicating that their initial tweet was personal mental health-related), and the remaining 123 users were labelled as: No. These datasets were not balanced as the ratio of the two classes was 65:35, hence we used stratification while splitting the dataset into Train and Test to ensure that the ratio of the two classes was maintained and that the results would not be skewed. The datasets were also shuffled before using stratification to split into 70% Train and 30% Test datasets.

4.3.2. Support vector machines

Support vector machines (SVMs) are a supervised learning model used for both classification and regression. The basic idea behind the SVM is to find a hyperplane that separates the classes in a feature space. In a two-dimensional space, the hyperplane is a line. The goal of SVM is to find the hyperplane which has the biggest gap or margin between the classes of data, also known as the Maximal Margin classifier. While it was originally designed for separating classes of data that were linearly separable, it is possible to efficiently perform non-linear classification using the kernel trick. Since the 1990s, using SVM for classification problems has grown mainly due to it's ease of use and high accuracy.

4.3.3. Deep learning using keras

Deep learning is an approach to machine learning inspired by our knowledge of neural networks in the human brain. Although artificial neural networks have been around for many years, deep learning has been recently growing in popularity especially due to the rise in availability of high computing power and the high increase in the volume of data. In essence, deep learning is nothing but a large artificial neural

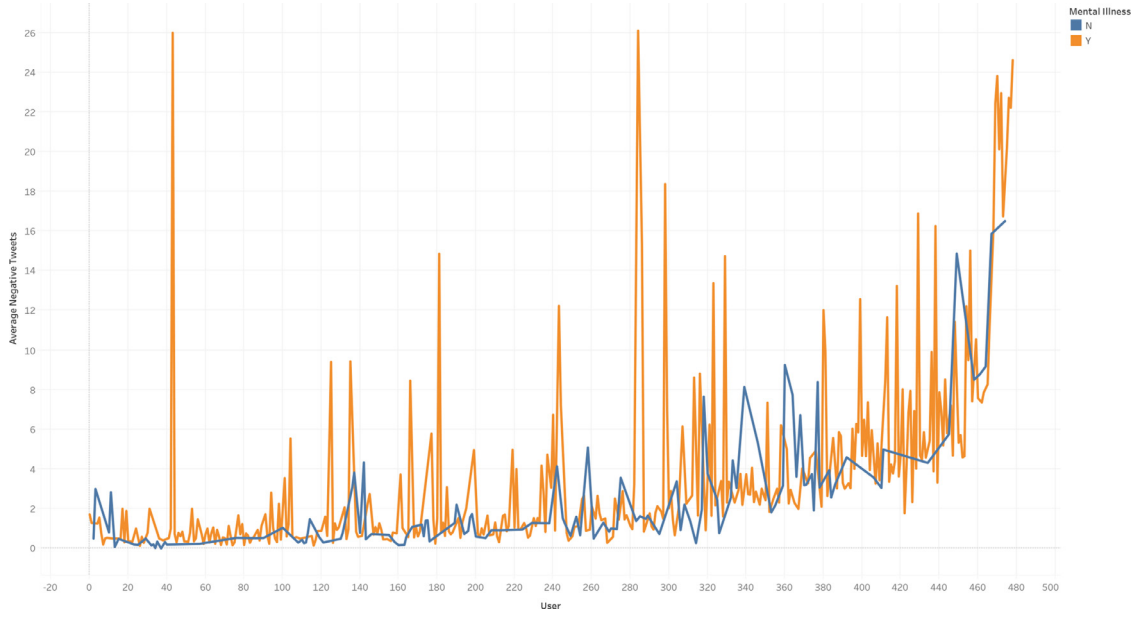

Fig. 3. Average number of negative tweets per user by class.

network, however, it is referred to as "deep" because of the approach to using many hidden layers of artificial neurons.

"Recurrent nets are a type of artificial neural network designed to recognize patterns in sequences of data, such as text, genomes, handwriting, the spoken word, or numerical times series data emanating from sensors, stock markets and government agencies. They are arguably the most powerful and useful type of neural network, applicable even to images, which can be decomposed into a series of patches and treated as a sequence." (Team, 2022).

We decided to explore Long Short-Term Memory (LSTM), a form of a gated recurrent neural network to help classify sequences of the number of negative sentiments by week over time by a user as likely to be suffering from mental illness. The User Weekly Negativity labelled dataset was used for this sequence-based classification using LSTM. This dataset was a 479×27 matrix (479 users, 27 weeks), excluding the class label column.

TensorFlow (Abadi et al., 2015) is an open-source software library for numerical computation using data flow graphs, originally developed by researchers and engineers working on the Google Brain Team for conducting machine learning and deep neural network research. We used the Keras (Chollet et al., 2015) library, a high-level neural network API that can run on top of TensorFlow. Keras is a user-friendly, modular and highly extensible library, making it fast for experimentation and prototyping.

We experimented with the following neural network configurations:

i Single Dense Layer (without LSTM) containing 108 hidden neurons
ii Single Layer LSTM with 108 hidden neurons and no dropouts
iii Single Layer LSTM with 108 hidden neurons and dropouts=0.2 between the hidden layer and output layer
iv Double Layer LSTM with 108 hidden neurons in first layer and 54 hidden neurons in second layer and no dropouts
v Double Layer LSTM with 108 hidden neurons in first layer and 54 hidden neurons in second layer and dropouts=0.2 between hidden layers and output layer

For all of the above configurations, the loss function was "binary cross entropy", the optimizer function was "adam", the activation function in the hidden layers was "relu", and the activation function at the output layer was "sigmoid". The number of epochs used was 100. Based on this setup, the performance of the proposed LSTM-based classification is evaluated. The results found from the simulations are discussed in the next section.

5. Results discussion

Sentiment analysis is hard and may not always be accurate because of the language usage, sarcasm and local lingo used based on culture or region. While we humans can distinguish sarcasm easily, it is still hard for even artificial neural networks to detect it and give an accurate sentiment score. Also, the local lingo used by people from different cultures or regions will generally have poor sentiment scores unless localized training datasets are used for sentiment classification.

The graphs for the average number of tweets by class (users labelled as mentally ill: Yes, No) and the average number of negative tweets by class are shown in Fig. 3. The graphs clearly indicate that the users labelled as mentally ill on an average tweet more and have a higher number of negative sentiment tweets.

We used SVM as our benchmark model on the User Summary dataset for classifying users as suffering from mental illness (Yes/No). We used the Random Forest classifier to rank the features by importance to select the top features from the dataset. The top two features were: The average Polarity and Negativity Ratio, which were used as input features to the SVM model.

We used sci-kit learns GridSearchCV to perform a 5-fold cross-validation on the Train dataset using SVM to find the best parameters. The best accuracy was 73.43% for parameters: C = 0.001, gamma = 0.0001, kernel = rbf, class weight = balanced. We explored Long Short-Term Memory (LSTM), a form of a gated recurrent neural network to help classify sequences of the number of negative sentiments by week over time by a user as likely to be suffering from mental illness.

The performance of the proposed LSTM based proactive mental health detection framework is evaluated based on the above mentioned prototype. The accuracy of the LSTM and the F-1 score is considered in this paper as an evaluation metric. It also compared the proposed mechanism with some benchmarks including SVM with ngrms (Abdelwahab, Bahgat, Lowrance, & Elmaghraby, 2015), semantic annotations (Saif, He, & Alani, 2012), ensemble feature engineering (Hassan, Abbasi, & Zeng, 2013) and random forest with pre-processing (Jianqiang & Xiaolin, 2017). The results are displayed below in Table 4. We observed that increasing the epochs resulted in higher training ac-

Table 4
Deep learning results.

Neural network type	Test accuracy
Proposed Single Dense Layer	**84.31**%
Proposed Single Layer LSTM	80.83%
Proposed Single Layer LSTM with Dropout	82.92%
Proposed Double Layer LSTM	78.06%
Proposed Double Layer LSTM with Dropout	80.83%
SVM with ngrms (Abdelwahab et al., 2015)	75.81%
Semantic Annotations (Saif et al., 2012)	72.87%
Ensemble Feature Engineering (Hassan et al., 2013)	81.89%
Random Forest with pre-processing (Jianqiang & Xiaolin, 2017)	81.21%

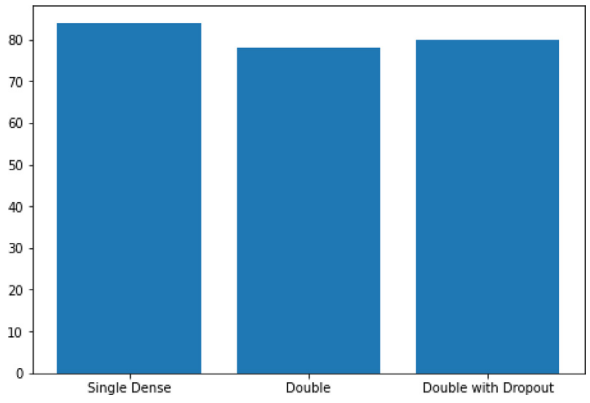

Fig. 4. F-1 score for varying number of layers in LSTM.

curacy but lower test accuracy which would be due to overfitting. The performance is also evaluated for F-1 score as shown in Fig. 4. For a varying number of layers in the LSTM, the F-1 score is calculated.

6. Contributions to literature and implications for practice

6.1. Contributions to literature compared to state-of-the-art

The advancement of machine learning and NLP approaches opens the door to unlimited opportunities for innovation in every aspect of our ecosystem (Mendhe, Henderson, Srivastava, & Mago, 2021). The key to improving the efficacy of machine learning and NLP services is to provide a large amount of data in the training process. Social media platforms such as Facebook, and Twitter are excellent hubs of large structured, unstructured, and meaningful data from human sensors. As a result, social media analytics has grown into a separate domain of knowledge (Guo, Yu, Li, Srivastava, & Lin, 2022). It is highly context-aware, i.e., different sections of society use social media analytics for different purposes. For example, the business community considers social media platforms as the centre of the business intelligence ecosystem that helps them better understand their consumers at a relatively lower cost (in contrast to costly survey-based approaches) (Huang, McIntosh, Sobolevsky, & Hung, 2017). The political community applies social media analytics to influence election results (Cheong & Cheong, 2011). The information ecosystem of social media produces fresh and live data which is an essential part of time-sensitive applications such as emergency response, live traffic analysis, etc. In this paper, we focus on a specific angle of social media analytics, i.e., sentiment analysis for health monitoring.

Sentiment analysis on Twitter is a topical research issue. Sentiment analysis and topic modelling on the tweets about COVID-19 in Singapore is performed to understand the public sentiment toward the COVID-19 outbreak in Singapore. Temporal topic trends extracted from Twitter and topic modelling are applied to find correlations between real-life events and sentiment changes throughout the whole period of lockdown in Singapore (Mohamed Ridhwan & Hargreaves, 2021). Machine learning techniques such as Naive Bayes, Decision Trees, Random Forests, and Support Vector Machines are applied for sentiment analysis, and classification of Indian farmers' protests using Twitter data (Neogi, Garg, Mishra, & Dwivedi, 2021).

A wide range of features and methods for training sentiment classifiers for Twitter datasets have been researched in recent years with varying results (Agarwal, Xie, Vovsha, Rambow, & Passonneau, 2011). Twitter sentiment analysis is difficult compared to general sentiment analysis due to the presence of slang words and misspellings. Both document and phrase-level sentiment identification approaches are proposed (Agarwal et al., 2011). Tree kernel is explored to obviate the need for tedious feature engineering from tweets. Semantic concepts are added to update the feature model with newer data and emoticons. The semantic features are being used to classify the polarity of the sentiments (Saif et al., 2012).

Deep learning on sentiment analysis evolved from the study of Artificial Neural Networks, improving upon them to fix issues with overfitting and training time (Bickman, 2020; Chakraborty & Kar, 2017; Graham et al., 2019; Neethu & Rajasree, 2013). The success of deep learning lies in a large number of hidden computation layers (Severyn & Moschitti, 2015). The DNN model is taught by being asked to predict the outcome of polarity classification based on a set of input parameters or features. The inference is then scored on how close it is to the real outcome, and that "score" or "error" is then propagated backwards through the model, updating the hyper parameters in the hidden processing layers referred to as gradient descent (Severyn & Moschitti, 2015). This directs the model to guess more correct results in the future. This process of gradient descent is repeated for the entire training dataset. If done correctly and with enough data, the model will be able to accurately predict the outcome of situations not included in the training dataset. An influence probability model for Twitter sentiment analysis is proposed in (Wu & Ren, 2011). A Multinomial Naive Bayes classifier that uses N-gram and POS-tags are used as features in training in Abdelwahab et al. (2015). Features are combined in an ensemble, and SVM is used as a base classifier in Hassan et al. (2013).

Deep-learning-based architectures for multi-modal sentiment classification are discussed in Garcia-Ceja et al. (2018); Poria et al. (2018). To boost the sentiment classification performance, a multi-modal data augmentation framework is offered in Xu, Mao, Wei, & Zeng (2020). Existing methods for distilling people's sentiments from the ever-growing amount of online social data are summarized in Cambria, Das, Bandyopadhyay, & Feraco (2017). Later on, Top-down and bottom-up learning are integrated via an ensemble of symbolic and sub-symbolic AI tools to detect the polarity from the text in Cambria, Li, Xing, Poria, & Kwok (2020). Notably, these approaches do not focus on mental health as significant pre-processing and semantics should be cured from the correlation of mental health and corresponding tweets. Moreover, we consider the sequential effects of successive tweets as it is usually observed in depressing tweets. Hence, existing approaches are not applicable as we focus on aggregating tweets or profiling, whereas traditional approaches analyse tweets individually.

Recently, there has been researched done on detecting mental illness based on a person's messaging and engagement on social media platforms (Banerjee & Shaikh, 2021; Benton, Mitchell, & Hovy, 2017; Karmegam, Ramamoorthy, & Mappillairajan, 2020). For example, building statistical classifiers to estimate the risk of depression based on users' behavioural attributes relating to their social engagement, emotions, linguistic style, social engagement and mentions of mental health-related medications (Choudhury, Gamon, Counts, & Horvitz, 2013). Another recent work (Reece et al., 2017), applied traditional machine learning classifiers to predict the emergence of depression and Post-Traumatic Stress Disorder (PTSD) in Twitter users who had been medically diagnosed as having PTSD or depression.

We have taken a different approach by collecting past tweets of users who initially mentioned that they had a mental illness such as depression. Our framework and prototype mainly focused on detailed

and custom data cleansing and performing sentiment analysis on historical tweets to predict the likelihood of a user who has a mental illness. Also, rather than only using traditional statistical and machine learning classifiers explored, we explored both traditional machine learning algorithms such as SVM for binary classification and the latest deep learning algorithms such as LSTMs for performing sequence-based classification to predict the likelihood of a user suffering from mental illness. We performed sentiment analysis using TextBlob and applied machine learning algorithms such as SVM and compared them with deep learning techniques using Keras on TensorFlow. The results show that traditional machine learning techniques are still efficient and accurate as compared to the new deep learning techniques. Consequently, our proposed approach contributes to real-time information management surrounding mental health.

6.2. Implications for practice

The major implication for practice of the proposed mechanism is the user-driven data collection. Patients are not always comfortable with regular sharing of the mental health status. Some patients are technologically challenged on sharing data. If the data collected continues to grow over time, deep learning techniques such as LSTM will become more efficient for classification.

The tools used in this prototype are scalable and deployable on a cloud-based platform for a large-scale production-based system. This work can be extended in future, considering the neutrality and ambivalence during the classification. Collecting reliable data is another implication for menatl health moniroting in pro-active way. There exists a large amount of fake users in the social media. Diagonosis and identifying reliable data source is another scope of this research.

7. Conclusion

This paper focused on detecting mental illness on social media platforms such as Twitter. Psychiatrists or mental health professionals can use such a framework or methodology to better understand their patients' sentiments. Governments and health agencies can use it to better understand the regions where people are suffering from such illnesses and provide necessary services. It can also be applied to any other social behaviour detection such as cyber-bullying, monitoring employee health in the workplace and more.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., & Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org, https://www.tensorflow.org/.

Abdelwahab, O., Bahgat, M., Lowrance, C. J., & Elmaghraby, A. (2015). Effect of training set size on SVM and naive Bayesfor twitter sentiment analysis. In *2015 IEEE international symposium on signal processing and information technology (ISSPIT)* (pp. 46–51). IEEE.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). Sentiment analysis of twitter data. In *Proceedings of the workshop on language in social media (LSM 2011)* (pp. 30–38).

Agarwal, N., Chauhan, S., Kar, A. K., & Goyal, S. (2017). Role of human behaviour attributes in mobile crowd sensing: a systematic literature review. *Digital Policy, Regulation and Governance, 19*(2), 168–185. 10.1108/DPRG-05-2016-0023.

Banerjee, S., & Shaikh, N. F. (2021). A survey on mental health monitoring system via social media data using deep learning framework. In *Techno-societal 2020* (pp. 879–887). Springer.

Batty, G. D., Russ, T. C., Stamatakis, E., & Kivimäki, M. (2017). Psychological distress in relation to site specific cancer mortality: Pooling of unpublished data from 16 prospective cohort studies. *British Medical Journal, 356*. 10.1136/bmj.j108.

Beal, V. (2016). Twitter dictionary: A guide to understanding twitter lingo. https://www.webopedia.com/quick_ref/Twitter_Dictionary_Guide.asp/.

Benton, A., Mitchell, M., & Hovy, D. (2017). Multi-task learning for mental health using social media text. arXiv preprint arXiv:1712.03538

Bickman, L. (2020). Improving mental health services: A 50-year journey from randomized experiments to artificial intelligence and precision mental health. *Administration and Policy in Mental Health and Mental Health Services Research, 47*(5), 795–843.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis* (pp. 1–10). Springer.

Cambria, E., Li, Y., Xing, F. Z., Poria, S., & Kwok, K. (2020). Senticnet 6: Ensemble application of symbolic and subsymbolic ai for sentiment analysis. In *Proceedings of the 29th ACM international conference on information & knowledge management* (pp. 105–114).

Chakraborty, A., & Kar, A. K. (2017). Swarm Intelligence: A Review of Algorithms. In S. Patnaik, X. S. Yang, & K. Nakamatsu (Eds.). *Nature-Inspired Computing and Optimization. Modeling and Optimization in Science and Technologies*: 10. Cham: Springer 47–494. 10.1007/978-3-319-50920-4_19.

Chen, M., Shen, K., Wang, R., Miao, Y., Jiang, Y., Hwang, K., ... Liu, Z. (2022). Negative information measurement at ai edge: A new perspective for mental health monitoring. *ACM Transactions on Internet Technology (TOIT), 22*(3), 1–16.

Cheong, F., & Cheong, C. (2011). Social media data mining: A social network analysis of tweets during the Australian 2010–2011 Australian floods. In *Pacis 2011-15th pacific asia conference on information systems: Quality research in pacific* (pp. 1–16). RMIT University.

Chollet, F. et al. (2015). Keras. https://github.com/fchollet/keras.

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Icwsm*.

Drydakis, N. (2021). Mobile applications aiming to facilitate immigrants' societal integration and overall level of integration, health and mental health. Does artificial intelligence enhance outcomes? *Computers in Human Behavior, 117*, 106661.

Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., & Tørresen, J. (2018). Mental health monitoring with multimodal sensing and machine learning: A survey. *Pervasive and Mobile Computing, 51*, 1–26.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports, 21*(11), 1–18.

Guo, Z., Yu, K., Li, Y., Srivastava, G., & Lin, J. C.-W. (2022). Deep Learning-Embedded Social Internet of Things for Ambiguity-Aware Social Recommendations,. *IEEE Transactions on Network Science and Engineering, 9*(3), 1067–1081 1 May-June 2022. 10.1109/TNSE.2021.3049262.

Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management, 42*, 78–89.

Hassan, A., Abbasi, A., & Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In *2013 international conference on social computing* (pp. 357–364). IEEE.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.

Huang, S.-C., McIntosh, S., Sobolevsky, S., & Hung, P. C. (2017). Big data analytics and business intelligence in industry. *Information Systems Frontiers, 19*(6), 1229–1232.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., & Huang, Z. (2020). Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems, 8*(1), 214–226.

Jianqiang, Z., & Xiaolin, G. (2017). Comparison research on text pre-processing methods on twitter sentiment analysis. *IEEE Access, 5*, 2870–2879.

Karmegam, D., Ramamoorthy, T., & Mappillairajan, B. (2020). A systematic review of techniques employed for determining mental health using social media in psychological surveillance during disasters. *Disaster Medicine and Public Health Preparedness, 14*(2), 265–272.

Kemp, S. (2017). *Digital in 2017: Global overview*. DGT Magazine. https://wearesocial.com/special-reports/digital-in-2017-global-overview

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports, 10*(1), 1–6.

Kouzis-Loukas, D. (2016). Learning scrapy.

Lab, M. N. (2017). Natural language toolkit (nltk project). http://www.nltk.org/.

Lee, K., Agrawal, A., & Choudhary, A. (2013). *Real-time disease surveillance using twitter data: Demonstration on flu and cancer* (vol. Part F128815, pp. 1474–1477). Association for Computing Machinery.

Marquez, P., & Dutta, S. (2017). Mental health. *Expert Systems*. http://www.worldbank.org/en/topic/health/brief/mental-health

Mendhe, C. H., Henderson, N., Srivastava, G., & Mago, V. (2021). A scalable platform to collect, store, visualize, and analyze big data in real time. *IEEE Transactions on Computational Social Systems, 8*(1), 260–269. 10.1109/TCSS.2020.2995497.

Mohamed Ridhwan, K., & Hargreaves, C. A. (2021). Leveraging twitter data to understand public sentiment for the COVID-19 outbreak in Singapore. *International Journal of Information Management Data Insights, 1*(2), 100021. 10.1016/j.jjimei.2021.100021.

MongoDB, I. (2018). mongodb. https://www.mongodb.com/.

Neethu, M., & Rajasree, R. (2013). Sentiment analysis in twitter using machine learning techniques. In *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)* (pp. 1–5). IEEE.

Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights, 1*(2), 100019. 10.1016/j.jjimei.2021.100019.

pattern.en. (2017). https://www.clips.uantwerpen.be/pages/pattern-en.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Gelbukh, A., & Hussain, A. (2018). Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems, 33*(6), 17–25.

Ranna Parekh, M. M. D. (2015). What is mental illness? https://www.psychiatry.org/patients-families/what-is-mental-illness.

Reece, A. G., Reagan, A. J., Lix, K. L. M., Dodds, P. S., Danforth, C. M., & Langer, E. J. (2017). Forecasting the onset and course of mental illness with twitter data. *Scientific reports*.

Saif, H., He, Y., & Alani, H. (2012). Semantic sentiment analysis of twitter. In *International semantic web conference* (pp. 508–524). Springer.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web, WWW '10* (pp. 851–860). New York, NY, USA: ACM.

Severyn, A., & Moschitti, A. (2015). Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval* (pp. 959–962).

Sharma, A., Sanghvi, K., & Churi, P. (2022). The impact of instagram on young adult's social comparison, colourism and mental health: Indian perspective. *International Journal of Information Management Data Insights, 2*(1), 100057.

Sims, S. (2017). 50 acronyms social media professionals must know. https://socialmediaweek.org/blog/2017/01/must-know-social-media-acronyms/.

Steven Loria. Textblob: Simplified text processing. http://textblob.readthedocs.io/en/dev/.

E. D. D. Team Deeplearning4j: Open-source distributed deep learning for the JVM, apache software foundation license 2.0. http://deeplearning4j.org.

Twitter, I.. Advanced search. https://twitter.com/search-advanced?lang=en.

Twitter, I. (2018). Twitter developer api. https://developer.twitter.com/.

Valdivia, A., Luzón, M. V., Cambria, E., & Herrera, F. (2018). Consensus vote models for detecting and filtering neutrality in sentiment analysis. *Information Fusion, 44*, 126–135.

Wikimedia Foundation, I.. Twitter. https://en.wikipedia.org/wiki/Twitter.

Wu, Y., & Ren, F. (2011). Learning sentimental influence in twitter. In *2011 international conference on future computer sciences and application* (pp. 119–122). IEEE.

Xu, N., Mao, W., Wei, P., & Zeng, D. (2020). MDA: Multimodal data augmentation framework for boosting performance on sentiment/emotion classification tasks. *IEEE Intelligent Systems, 36*(6), 3–12.

Zhu, J., Fang, F., Sjölander, A., Fall, K., Adami, H. O., & Valdimarsdóttir, U. (2017). First-onset mental disorders after cancer diagnosis and cancer-specific mortality: A nationwide cohort study. *Annals of Oncology, 28*(8), 1964–1969. 10.1093/annonc/mdx265.