Mental Anxiety and Depression Detection during Pandemic using Machine Learning

[1]Quadir, Ryana & [2]Hossain, Md. Fokhray

Department of Computer Science & Engineering

Daffodil International University

Email: ryana25-866@diu.edu.bd, drfokhray@daffodilvarsity.edu.bd

Abstract:

With the rise of Social Media usage, web surfing and a long period of uncertainty during the Pandemic time, there is a sheer concern about the mental health and anxiety disorders among people. This increate rate of online social media use (SMU) opened the possibility to identify some common traits among people with various mental disorders and anxiety by the large dataset provided. In recent years, this research area has started to evolve, but it would be extremely valuable during this crisis period. Although it is a complex task to perform as mental illness patterns are very complicated, it showed the light of hope in the past. Previously, adoptive supervised machine learning, such as deep neural network approaches were used to predict the pattern and level of mental illness; but they failed due to lack of annotated training data. In this paper, we are proposing an effective machine learning architecture, based on Cluster analysis and Natural Language Processing (NLP) technique in the analysis of unstructured data extraction from Social media platforms.

Keywords:
Mental Anxiety, Social media, SMU, Depression, Machine learning, Cluster analysis

1. INTRODUCTION

With the outburst of the coronavirus disease Covid-19 early in 2020, the time spent at home has increased sharply for people around the world. But it was after the upsurge of the pandemic, somewhere in between February and March, people started to experience strict lockdown situations to help flatten the pandemic's deadly curve. During a long stay at home, they tend to spend more time on Social Media platforms, providing an unexpected boost to engagement on these platforms. At the same time, as a byproduct of the coronavirus disease, people start to feel a strange kind of stress, fear, and anxiety about being put in such an area of uncertainty. Even before the pandemic, 1 out of 13 people globally suffers from anxiety, according to the World Health Organization (WHO). The WHO reports that anxiety disorders are the most common mental disorders worldwide with specific phobia, major depressive disorder and social phobia being the most common anxiety disorders. For the current situation, of course there is grief about the sufferings and deaths that people are witnessing everyday; but even so, the helplessness is even stronger when people feel they are losing things in terms of the ability to move around freely. These negative feelings are reflected directly on people's daily activities and consequently exposed in Social Media. It is not just in terms of sharing the stress, frustration or fear, it is sometimes only the connection and impact of the overall Social Media. In other words, the association between SMU and mental health may be indicative of a user's experience and attitude rather than the volume of SMU. In this regard, SMU would only be used as a measure to detect the users' mental stability and psychiatric disorder (if there is any); but not to demonstrate the negative effect of Social Media usage among them.

2. RELATED WORK

There were many approaches made for detecting mental disorder and depression of the social media users in the past. Many of them dealt with the volume of usage as a measure of detecting potential SNMD (Social Network Mental Disorder). That means, they had a different goal to prove than ours. Nonetheless, these models helped us identify the feature extraction well and to formulate a better classification for clusters.

2.1: Affective Content Analysis of Online Depression Community by Thin Nguyen

Thin Nguyen et al. used affective content analysis method of Online Depression Community to design the algorithm for detecting depression stigma. The dataset was collected from Weibo and Tweeter- two of the well-known social media platforms. In this process, first, the collected dataset was analyzed using syntax and semantic analysis, which gave the sense of depression stigma among the social media posts by different age groups. The posts where keywords matched but didn't qualify for stigma-set, belong to the Non-stigma group. The posts were then classified according to the depression and mental disorder symptoms. The data-modeling performance was improved by implementing classification algorithm like, SLR (Simple Logistic Regression), SVM (Support Vector Machine), RF (Random Forest) etc.

This approach only worked with users' social media posts and their syntactic analysis and classification; but there could be some feelings indicating depression stigma that were not exposed in the posts. That area wasn't dealt in this process.

2.2: Depressive Symptoms Analysis using Active and Passive SMU

Social Media Usage was classified as Active and Passive to analyze depression symptoms by César G. Escobar-Viera, Ariel Shensa, Jaime E Sidani, Brian A. Primack, Nicholas David Bowman, Jennifer Knight and Everette James. Their classification was mostly based on the way social media is used, how much time people are spending on it and how many platforms are they visiting. The Active users are found out to be of improved well-being whereas the Passive users are likely to have social anxiety and of decreased well-being. These opposing effects had varied association with depression symptoms. They examined their data-distribution of 7 SMU variables using Shapiro-Wilk test of Normality and histogram, Q-Q plots graphical methods. They initiated the process from the hypothesis that passive users would have positive association with mental illness and depression.

This straightforward association between usage pattern and depression turned out to be true that depression was not found in the active users, which proved their initial hypothesis. But active use of social media can sometimes involve problematic SMU and addiction, which is a potential indicator of mental illness. That extreme case should have been considered in constructing the model.

2.3: Tensor techniques applied in Multiple Online Social Networks

Preserving users' privacy with the help of separating agents, Sridharan et al. proposed the depression detection diagnostics using Convolution Neural Networks (CNN). As the dataset, he used the data posted by different users in social networks. Another filtering mechanism was used to filter out redundant data to increase the learning of CNN. This approach made an attempt to automatically identify potential online users with Social-Network online disorder detection-framework. It managed to explore various features from various data-logs and this is where a new tensor technique was applied for deriving latent features from multiple Social Networks.

Again, this method worked only with the posts from users and tried to identify the latent features from there. But the intensity of social media usage or emotional attachment pattern were not dealt in finding the disorder or depression symptoms.

3. METHOD

As we stated before, it is a complicated task to determine the mental illness as the disorder patterns are very unpredictable and complicated, we developed patterns of different user-clusters with extreme vulnerability. For example, one cluster might have users with addictive or problematic levels of SMU, which may be associated with increased anxiety and mental stress. This association might be placeable to the increased likelihood of individuals who experience long term anxiety specially during pandemic and subsequently develop addictive behaviors. On the other hand, there is a group of sensitive users who do moderate level of SMU but has immensely strong emotional connection with social media. So, we would like to define these clusters first and then start to ascertain the susceptibility to severe anxiety and alarming level of depression.

Besides, mental disorder is perceived not by observing someone feeling down or depressed or fearful just for a moment; it would persist for longer time instances that affects general cognitive function of the brain. And as a consequence of witnessing the death toll around and the inability to interact socially, the mental illness became a common phenomenon worldwide.

As we did not specify how many clusters there would be, we chose the Divisive Hierarchical Clustering in this regard. This clustering technique allowed us to identify optimal number of clusters with distinguishing usage-pattern. After feeding data against all 3 parameters, our IDE: RStudio unpacked the packages `factoMineR` and `factoExtra` and with the help of package `ggplot2` and `graphics`, we concluded with a Dendrogram-plot with 4 distinctive clusters.

While choosing social media platforms, we selected 10 most widely used sites, namely: Facebook, YouTube, Instagram, Twitter, Tiktok, Weibo, Snapchat, Pinterest, QQ, and Qzone. Now, considering the time spent on these social media, frequency of usage, intensity of use and level of abusing the SMU, these are the 4 clusters:

1. Low Usage (LU): Not affecting personal life
2. Moderate Usage (MU): Strong emotional connection
3. High Usage (HU): High but not problematic
4. Addictive Usage (AU): Problematic and addictive social media uses

The measurable parameters that are used to define these clusters are:

i. Time spent (per day)
ii. Frequency of use (per day)
iii. Number of SM platforms

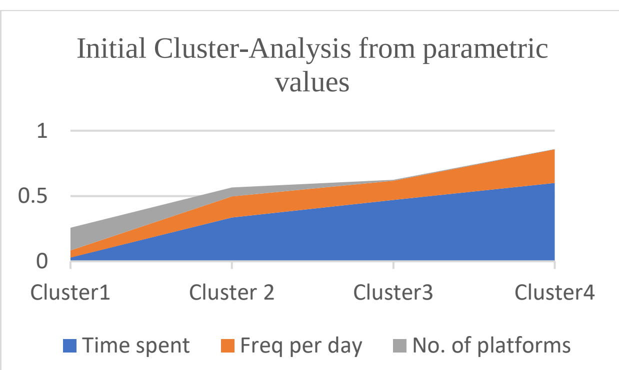
First, an initial classification was drawn for 4 clusters based on these parameters above. Here, we used the data from our social network and online statistics from datareportal.com[3], businessinsider.com[4], statista.com[5], smartInsights.com[6], globalWebIndex.com, jmir.org[7] tables. Total 200 respondents from our network provided their statistics about average spending time daily on social media, number of times these platforms visited per day and number of social media active accounts they have and use during COVID-19 period. About 100 of them also participated in filling up tables for their emotional statistics. The ranges for each parameter were specified after analyzing the collected data by the Cluster analysis. After having the product of cluster range-value and their corresponding weightage, and by normalizing the data, the first table looked like this:

Table 1:

Parameters	Cluster1 LU	Cluster2 MU	Cluster3 HU	Cluster4 AU
Time spent	0.02835 (<1)	0.3366 (1-2)	0.4695 (2-3)	0.6 (>3)
Frequency	0.055 (1-3)	0.15963 (7-9)	0.14852 (4-6)	0.2575 (>9)
No. of platforms	0.175 (1-4)	0.06875 (4-7)	0.0056 (7-9)	0.00211 (>9)

Table1: Cluster degree of memberships by social media usage parameters

After drawing the initial clusters, the graph looked like this:



Secondly, we gathered posts and popular hashtags of these individuals and started to extract the keywords using NLP (Natural Language Processing) from these posts for syntactic analysis. Again, we relied on R and package was mostly `Udpipe` and dependencies were `dplyr` and again `ggplot2`. Udpipe Package provides pretrained language models that we used.

Here, we collected mostly Facebook, Twitter and Weibo posts and used NLP to identify the most focused words and examined the effective keywords by tuning them up. After getting the keywords, the association was made from keyword-based analysis regarding individual’s mental condition.

So, from 2nd phase, we got our refined clusters with criteria-related validity and 3 emotional indices from Kaggle Andbrain_dataset. A part of the analysis looked like this:

Table 2:

SL	Posts and hashtags	Keywords	Tuned up keyword	Depression-index	Anxiety-index	Fear-index
1	97,000 kids tested Covid-19 positive in the US within 2 weeks of School opening. Bangladesh is planning to open schools by September.	Tested, Covid-19, positive, planning, open	Covid-19, positive	0.2114	0.5168	0.358
2	Empty airports and planes at the height of the #COVID-19 pandemic	Empty airports, height, pandemic, planes	Empty, pandemic	0.41	0.35	0.112
3	Encouraging people to get #Coronavirus in the name of Virtue	virtue, people	virtue, people	0.516	0.131	0.082
4	This implies there will be no 'after' of it and if there's no vaccine, this virus will remain with us	virus, vaccine	virus, vaccine	0.423	0.315	0.259
5	Trying not to be angry about this #coronavirus is hard, talking listening and even seeing the effects of it	angry, effects	angry, effects	0.457	0.241	0.238

Table 2: Keyword extraction from posts and analyzed with emotion-index

Thirdly, we tried to get the users' emotional statistics and for that we asked 100 participants from all 4 clusters of the same sample to complete the table about their feelings over the last 7 days. The PROMIS (Patient-Reported Outcomes Measurement Information System) depression-scale is a well stablished way of obtaining the emotional state of people. This is a 4-item scale of measuring users' feelings and items are scored using a 5-point Likert-type scale. It was applied to all the clusters including Cluster1 (Low Usage). This step is for emphasizing on the emotional attachment with social media even when users are not exposing themselves enough to the online world.

The raw scores from this table would have a range from 4 to 20 and greater scores indicate increased severity of mental issue, specially depression symptoms. The probabilities were calculated from individuals' score from this table:

Table 3:

	Emotions	Never (1)	Rarely (2)	Sometimes (3)	Often (4)	Always (5)
1	Hopeless			√		
2	Worthless		√			
3	Helpless				√	
4	Depressed			√		

Table 3: An example of how a participant completed the 4-item form

In the final phase, there were again 100 respondents who agreed on completing the 7-item Generalized Anxiety Disorders Scale (GAD-7) to have a clear identification of the anxiety vulnerable users among the clusters. GAD-7 is a screener for generalized anxiety disorder in primary care settings. Those respondents were asked to state the frequency of the 7 items that they have been bothered over the last 4 weeks. Again, the probability of having Anxiety-symptom was calculated for individual's responses. The GAD-7 looked like this:

Table 4:

	Emotions	Never (0)	Several days (1)	Over half the days (2)	Nearly everyday (3)
1	Feeling nervous, anxious, or on edge		√		
2	Not being able to stop worrying		√		
3	Worrying too much about different things				√
4	Trouble relaxing			√	
5	Being so restless that it's hard to sit still	√			
6	Becoming easily annoyed or irritable				√
7	Feeling afraid as if something awful might happen			√	

Table 4: An example of how a participant completed the GAD-7 form

The Anxiety measure here would be: (1+1+3+2+0+3+2)/21 = 12/21 = 0.5714

4. DATA ANALYSIS

4.1: From Table1, we can see the cluster-membership was classified by percent of users approving some levels of each SMU characteristic. Also, not any one cluster is exclusively giving high values for all parameters. It is rather 3 of them- MU,HU and AU which were seemingly found as vulnerable in terms of higher value. The collinearity between these predictor variables was assessed and Cluster1 (LU) fell below 0.01, that justifies the initial classification. But we are not going to omit this cluster for the next step and parsing would be done on posts of all 4 cluster users.

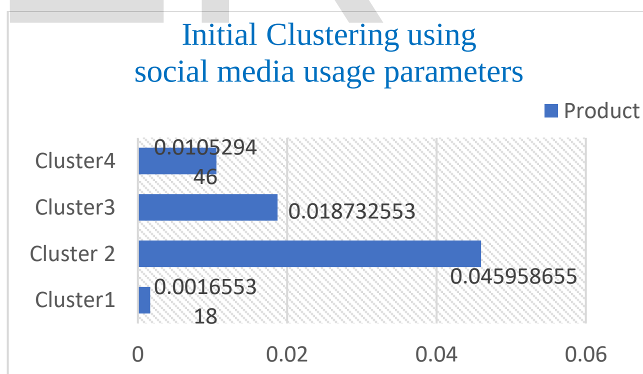


Fig1: Collinearity between Cluster-variables

4.2: Next, the keywords were obtained from the social media posts which were used as positive-labelled training data for an NLP classifier. Here, we took just one post from single user. For the sentiment analysis, these keywords were projected on emotions sensor dataset containing top English words classified statistically using Naive Bayes

Algorithm into 7 basic emotions. We picked up indices of three emotions that are relevant with our goal: fear, anxiety and depression. These indices are then examined for multi-collinearity by the multivariate correlation matrix. This helped evaluating the criterion-related validity of the cluster solution. At the end of this step, we'll have a more concise version of the correlation.

5	Depression-index	Anxiety-index	Fear-index
Depression-index	1 0.208849	0.110137	0.108766
Anxiety-index	0.110137	1 0.058081	0.057358
Fear-index	0.108766	0.057358	1 0.056644

Fig2: Multivariate correlation matrix

4.3: After step2, we'll get more improved classification of the clusters and identify those individuals having higher index-values for anxiety and depression and their belonging clusters. Now, we would analyze the PROMIS-score for users' depression measurement. For that we would examine the probability of each table filled up by users. Then the probabilities would be ranked on Depressive symptoms like:

Symptom level	Calculated Probability	Cluster
None to slight	P < 0.4	LU
Mild	0.4< P<=0.5	HU
Moderate	0.5< P =0.8	MU
Severe	0.8< P = 1	AU

Fig3: Depressive Symptom level from PROMIS

4.4: Lastly, we would focus on individual's anxiety and fear measurement during pandemic and the probability of anxiety level was calculated from the GAD-7 tables.

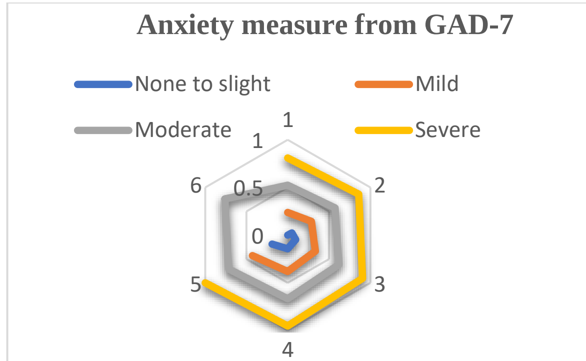


Fig4: Anxiety level from GAD-7 table

5. RESULT AND DISCUSSION

We started the entire procedure with 200 participants' statistics from our network and other stats from online portals for initial cluster analysis and analyzed 250 posts from social media platforms of 200 users. In the third phase, we managed to get 100 individuals who completed the PROMIS form and we ended up having the same 100 people filling up the final GAD-7 table. Although the number of participants along with the percentages varied in different phases, it yielded 4 distinct patterns of their social media use and relevant mental condition.

During the initial cluster analysis, we found that the LU-cluster (Low Usage) had lowest score for all the 3 variables- time, frequency and no. of platforms use. As a result, they had fewer participation in posting feelings in social media. For applying the 2 psychological disorder screeners, we did not take out the LU-cluster even though they failed to look vulnerable in terms of any mental ailment. From the keyword extraction phase, this LU-cluster began to fade out and ended up as the group that is 'not associated' with an any mental illness.

For the MU-cluster, the time spent was a moderate 1-2 hours and they had 4-7 active SM accounts, but the frequency is in bit higher range (7-9 times per day). This group indicates those people who spent not a huge amount of time but have strong emotional connection with social media and engage in attention seeking behaviors like frequent status-updates and subsequent checking for 'Like' in online world. This obsession may lead to depression if the individuals do not receive the desired feedback from his or her social media audience. It might get elevated during lockdown phase which was reflected from the PROMIS-scale result as they showed moderate (0.5-0.8) probability of Depressive symptoms. No wonder, they demonstrated maximum participation and hence had more posts for keyword extraction.

We found interesting analysis for HU-cluster, who spent 2-3 hours every day having 7-9 active SM accounts. They tend to stay connected to their online world and showed less frequency in checking their own account for feedback. In both psychological disorder screeners, they exhibited probabilities below 0.5, signifying their positive mental state. They tend to share life experience even during pandemic-time and respond quickly and frequently to other users which leads to their improved well-being.

Finally, we defined Cluster-4 as 'Addictive' since all 3 variables associated with them gave us the highest values in doing initial cluster-analysis. The 'Fear Of Missing Out' (FOMO) felling made them stay continually connected in social media losing real-life interactions triggered the elevated anxiety and fear symptoms. In case of keyword extraction, their posts indicated high index-values of not just depression and anxiety, they were also linked with fear. Additionally, they unveiled probabilities in the range 08-0.95 in both psychological tests which yields extremely high risk of mental syndromes.

Scores	0-4	5-10	11-16	17-21
Probability	0 - 0.1904	0.238 – 0.47619	0.5238-0.7679	0.809-0.9523
Cluster	LU	HU	MU	AU

Table 5: Results from GAD-7 scale

Hence from this 4-phase analysis, it was clear that people from Cluster-2 (MU) and Cluster-4 (AU) are closely associated with mental ailment.

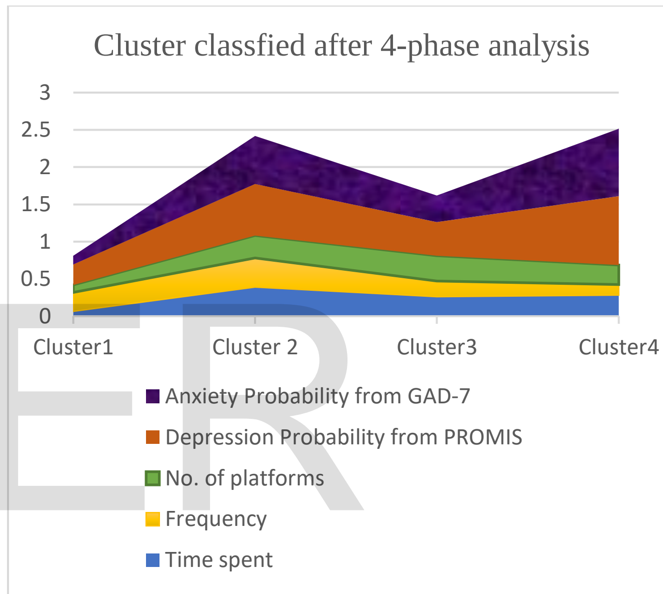


Fig5: Final Cluster result from 4-phase analysis

However, some border-line values of Cluster-3 (HU) individuals also need to have a close look at the mental health for their well-being.

6. LIMITATIONS

The dataset we constructed was entirely from our network and thus we got the values for the two psychological screeners. Without accessing our own network, these sensitive values were not possible to collect. On the contrary, these responses could not get diverse in terms of different Ethnicity and race and social positions. Yet, we tried to have variations in selecting social media users concerning their age, gender, occupation, and religion. We could have done an analysis of Phase-1 and phase-2 from online datasets, but the mental-health screening would not be conducted in that case. Even though this dataset is seemingly small, the data collected were authentic and legitimate and the information were taken carefully preserving their privacy.

7. CONCLUSION

The deterioration of people's mental health during pandemic period can result in severe mental disorder. For different people, the combination of emotions is different- for most of them, it's anxiety for the family, relatives and his/herself. It is a tremendous grief for losing our normal day-to-day social interactions which leads to severe depression. Sometimes it is not just general anxiety or depression, it is the fear about what is going to happen to the society, to the economy and for how long it is going to last. If the uncertainty lasts for more than a year and people suffer from that during the entire time or longer, it would become daunting. And it needs to be identified and treated as early as possible. Detecting earlier depression and anxiety can be a huge step to address the mental illness and offer support to the people suffering from this terrible mental illness. This four-phase approach using cluster analysis, NLP and two psychological screener that we proposed in this paper and the associations between the parameters are quite relevant and obtainable. Among the 4 distinct clusters, the 2 specific patterns- MU and AU with elevated symptom levels of depression and anxiety are undoubtedly at higher risk. This realistic diagnosis of mental disorder should be considered seriously in clinical interventions and therefore should be taken care at the early stage for the individuals at highest risk.

REFERENCES

1. Ce´sar G. Escobar-Viera, Ariel Shensa, Nicholas D. Bowman, Jaime E. Sidani, Jennifer Knight, Everette James and Brian A. Primack. 2018. Passive and Active Social Media Use and Depressive Symptoms Among United States Adults. In CYBERPSYCHOLOGY, BEHAVIOR, AND SOCIAL NETWORKING Volume 21, Number 7.
2. Devakunchari Ramalingam, Vaibhav Sharma and Priyanka Zar. May 2019. Study of Depression Analysis using Machine Learning Techniques. In International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075,Volume-8, Issue-7C2
3. Simon Kemp. January 30, 2020. Digital 2020: Global Digital Overview. Retrieved from DataReportal: https://datareportal.com/reports/digital-2020-global-digital-overview
4. Alexandra Samet. June 9, 2020, 10:58 PM. 2020 US SOCIAL MEDIA USAGE: How the Coronavirus is Changing Consumer Behavior. Retrieved from Business Insider URL: https://www.businessinsider.com/2020-us-social-media-usage-report?r=US&IR=T
5. J. Clement. June 19, 2020. Estimated U.S. social media usage increase due to coronavirus home isolation 2020. In reports from Statista URL: https://www.statista.com/statistics/1106343/social-usage-increase-due-to-coronavirus-home-usa/

6. Dave Chaffey. August 3, 2020. Global social media research summary July 2020. From smartInsights URL: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/
7. Ahmad AR, Murad HR. May 19, 2020. The Impact of Social Media on Panic During the COVID-19 Pandemic in Iraqi Kurdistan: Online Questionnaire Study. Vol 22, No 5 (2020). Retrieved from: https://www.jmir.org/2020/5/e19556/
8. Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi and Diana Inkpen. June 5, 2018. Deep Learning for Depression Detection of Twitter Users. In Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 88–97 New Orleans, Louisiana.
9. Bethany McCord, Thomas L. Rodebaugh, Cheri A. Levinson. 2014. Research Report named Facebook: Social uses and anxiety. Computers in Human Behavior 34 (2014) 23–27. Retrieved from www.elsevier.com/locate/comphumbeh
10. Margaret Mitchell, Kristy Hollingshead and Glen Coppersmith. 13 December 2016. Quantifying the Language of Schizophrenia in Social Media. Retrieved from https://www.researchgate.net/publication/291373346
11. Kirsten Weir. March 16, 2020. Seven crucial research findings that can help people deal with COVID-19. In American Psychological Association. Retrieved from URL: https://www.apa.org/news/apa/2020/03/covid-19-research-findings
12. Samantha K Brooks, Rebecca K Webster, Louise E Smith, Lisa Woodland, Prof Simon Wessely, Prof Neil Greenberg, et al. March 14, 2020. The psychological impact of quarantine and how to reduce it: rapid review of the evidence. Volume 395, ISSUE 10227, P912-920.
13. Katie Keane and Elise Ruland. In R2integrated blog. Retrieved from: https://www.r2integrated.com/r2insights/covid19s-impact-on-social-media-usage-and-strategies
14. Emily Chen, Kristina Lerman and Emilio Ferrara. Published online May 29, 2020. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. Retrieved from URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7265654/
15. C. O. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In Proceedings of HLT/EMNLP, pages 579–586
16. Sverre Urnes Johnson, Pål Gunnar Ulvenes, Tuva Øktedalen and Asle Hoffart. Published online Aug 6, 2019. Psychometric Properties of the General Anxiety Disorder 7-Item (GAD-7) Scale in a Heterogeneous Psychiatric Sample. In Front Psychol. 2019; 10: 1713. Retrieved from URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6691128/#:~:text=The%20seven%20items%20assess%20(1,if%20something%20awful%20might%20happen.