ISSN 2278-3091
Volume 11, No.1, January - February 2022
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse041112022.pdf
https://doi.org/10.30534/ijatcse/2022/041112022

To Analysis the Relation Between Internet Usage and Depression with Machine Learning

Amarjit Malhotra[1], Megha Gupta[2], Varun Gupta[1], Sanchit Shokeen[1], & Ankit Singla[1]
[1]Department of Information Technology, Netaji Subhas University of Technology, India
E-mail: [1]uppalz_amar@yahoo.com, varun.gupta1798@gmail.com,
sanchit.shokeen.ss@gmail.com, asingla590@gmail.com
[2]Department of Computer Science, MSCW, University of Delhi, India
Corresponding author: [2]meghabis@gmail.com

Received Date : December 10, 2021 Accepted Date :January 09, 2022 Published Date : February 06, 2022

ABSTRACT

Depression is a major disorder in the population of the 21st century. Previous studies have associated depression with internet usage and as the access to internet spreads through the developing countries depression can prove to be a challenging issue to combat. This paper proposes a new method to detect the presence of general depression among social media users using features extracted from their online habits including browsing, streaming, games and social media. The pertinent data has been collected from individuals primarily based in developing countries and applied various established supervised machine learning algorithms to predict the presence of general depression. Preliminary testing shows that the proposed system performs rather well and enables an easy method for keeping online mental health in check without compromising on privacy. The proposed work shows a clear positive correlation between social media usage and general depression highlighting the inimical effects of elevated usage.

Key words: Depression, social media, Machine Learning, Anxiety, Self-esteem, Online Activity

1. INTRODUCTION

Since its very inception, social media has slowly taken over all aspects of people's lives. An average Indian spends over 2.4 hours on social media in line with the global average of 2.5 hours [1]. This amounts to over 16.8 hours of weekly use which is almost as much as a part time job. Studies have shown that FOMO (Fear of missing out) drives users to continually reach out for their phones and limiting usage can help reduce anxiety and loneliness [2]. The sharp rise in social media usage can hence directly be linked to rising depression and developing countries are especially at risk with inadequate infrastructure and bleak healthcare. India, for instance has an acute shortage of doctors and medical staff with the density falling even below the World Health Organization's (WHO) critical shortage threshold [3].

Previous studies have shown that social media activity provides a new lens into patients' lives and thoughts shared in the form of posts and tweets can be used to predict depression even before its onset [4]. This is clear and conclusive link between online habits of users and their mental health. It has been previously shown that adolescents who invested more time in social media were more likely to develop a host of mental problems like poor sleep quality, anxiety, lower self-esteem and depression [5].

The various proven causes of depression like abuse, conflict, exposure to major events etc. can all be found on social media in a modern form. The continuous ongoing conflict and the profound spread of negative information in these spaces is also a cause for concern. The lack of enforceable age verification means that young kids are exposed to these risks as well. The problem gets compounded due to primary care physicians being unable to detect depression half the time [6].
Hypothesis 1 - Expected to be able to predict the presence of general depression amongst social media users from India using various machine learning techniques.
Hypothesis 2 - Expected to find an inverse correlation between increased and more invested social media usage and general mental health of individuals.

2.RELATED WORK

Concerns regarding the effects of social media on mental health of populations have been a cause of concern for re-searchers since the very inception of social media itself. More than 264 million people suffer from depression worldwide [7] and the numbers have been increasing in recent years.

Studies have already associated internet usage with mental health problems in the past. Various machine learning algorithms have also been employed for sentiment analysis using data from social media platforms [8]. This analysis over a period of time can be quite indicative of a person's mental health and machine learning models produce accurate results in such scenarios. Previous studies have also linked

online games like Player Unknown's Battlegrounds (PUBG) to mental disorders like Internet Gaming Disorder (IGD) and Generalized Anxiety Disorder (GAD) [9]. An inverse relationship between video game violence and mental health has firmly been established by similar studies.

Machine learning algorithms have also been applied to determine measures of suicidality using twitter data in the US population [10]. The study found that classification algorithms can deter-mine suicidal tendency in a person with a high degree of accuracy using their online twitter activity. Other methods that have been employed to analyse mental health in the social media era include text analysis, image analysis and social interaction graphs [11]. Previous studies have tried to link IA (Internet addiction) to depression in the long term and have established that populations addicted to the internet are more susceptible to depression [12].

While multiple studies have been conducted in this area, none have tried to link the usage patterns and online activity of people directly to mental disorders and depression. With a sharp spike in the number of such services in the past few years it is pivotal to study both their long- and short-term effects. In this work, a new method has been proposed to detect general depression using features extracted from online habits and the Geriatric depression scale (GDS) short form by Stanford [13].

3.METHODOLOGY

3.1Current Study

In the current work, five types of Social Media interactions have been included: Facebook, Instagram usage trends and preferences, involvement in tracking influencer and celebrity lives, daily average video streaming time and content categories streamed (from providers like Netflix, Amazon Prime, Disney+ Hotstar, Youtube and other similar platforms), digital gaming activity and thoughts on video game violence. It is also taken in consideration that time spent in reading and engaging in discussions on apps like Quora and Reddit and type of content consumed. This data was then used to extract relevant features which were used as inputs to the machine learning algorithms. To identify the class for a particular data point the Geriatric depression scale (GDS) short form by Stanford [13] was also filled up by the participants. Supervised machine learning algorithms and classification techniques were then employed to train the model to detect the possibility of depression amongst the participants of the study. The algorithms used include "Logistic Regression", "Support Vector Machine", "Decision Tree", "K-Nearest Neighbor", "Multilayer Perceptron Artificial Neural Network", "Naive Bayes classifier" and "Random Forest".

3.2 Participants

The collection of Social Media usage statistics from educated Urban Indian citizens, falling in the age group of 15 to 30 has been done. India has 718 million Internet users as of 2019 [14], and is the second largest online market, behind China. Millennials and Generation Z combined comprise more than 80% of Social Media users in India [15]. Social Media usage, and consequently social media induced problems are therefore more prominent in the said age group, making it all the more suitable for this study.The data was accumulated by circulating the link to a Google Form, which was to be filled by the participants online. It was specifically communicated that the data from the form would be used to conduct a study on social media habits of participants. The form collected demographic information, presented questions relating to social media usage activity, statistics, and preferences of the users. All questions were required to be answered for successful submission of the survey. The form clearly indicated that the questions belong to "Social Media Habits" of the participant. For this study no consent has been taken from any board or committee. This is because of the non-clinical, non-medicinal, and non-intrusive nature of this work. Also, the authors are affiliated to a technology university that has no internal committee related to research on human subjects.

3.3 Data Collection

The survey was shared among students at various schools and colleges, through message groups and peer to peer communication. In order to encompass a larger audience, the survey was also shared through social media as well as in other Indian universities and colleges' bulletins. The form was composed in the English language, as it is widely spoken and recognized in India as well as the rest of the world.For this study, responses from 300 unique participants have been taken. On inspection, the data thus collected was diverse in both demo-graphics and responses. The participants included 185 males, 111 females, and 4 others who chose not to disclose their pronouns. 258 participants (86%) belonged to the age range of 18-25, while the other participants were almost equally distributed on both sides of this range. Data was collected in the period beginning February 2nd, 2020 and ending February 9th, 2020.

3.4 Pre-processing

Data collected with the help of the survey was mapped to an appropriate set of features, which were used to train several supervised machine learning classifiers that could detect the presence of depression in a said social media user. The demographic data however, did not play any part in this process.Questions to which a single option was to be selected as a response were mapped to a single feature while the questions which allowed multiple choices to be selected were mapped such that every option to the

question was taken as a binary feature. These questions asked participants to choose categories that were relevant to their streaming, reading/writing preferences, and their involvement in tracking influencer and celebrity lives on social media platforms. Data pre-processing was carried out in a Python 3 environment, using Sublime Text 3 code editor and the Pandas library to perform the said operations.

3.5 Feature Extraction

Following the extraction of said features from the results of this survey, it has been attempted to study the impact each one of these makes on the mental health of social media users. Some of these features produce a positive impact; however, most of them affect users' mental health negatively.Broadly, these features cover the following five ways in which users generally participate in social media activities:

1.Facebook, Instagram usage statistics and preferences - Users scroll their news feed, share posts, upload pictures on Facebook and Instagram. They may also buy products being advertised on these platforms. In addition to all these, features corresponding to whether or not a user checks a picture for likes after uploading one and how much do they think it affects them has been introduced. In addition, it has been also noticed that how users feel about the number of followers they have on Instagram. In total, these constitute 6 of the total 30 features.

2.Involvement in tracking influencer and celebrity lives -features pertain to determining the kind of involvement a user prefers to keep in following celebrities. The achieved results found that some users choose to not follow any celebrities, while some read about them in the news, follow them on social media, and some do not wish to miss out on anything about their favourite celebrities.

3.Average hours spent on streaming video and type of content streamed - These features capture information regarding the time that users are investing in streaming video content from various categories online, and genres they like to watch. Options included Sports, Comedy, Politics, Fantasy, Violence, News, Romance and Informational. Ten features to video streaming has been allocated.

4.Digital gaming involvement and thoughts on video game violence – It has been asked to the users about their views on video game violence and whether or not they liked to play such games. It has been chosen as a feature for this study since previous studies have linked violence in video games to probable mental disorders in both children as well as adults [9].

5.Time spent reading and engaging in discussions on Quora, Reddit under several categories - 50% of the features in this category correspond to whether or not participants like to use websites like Quora and Reddit and whether they use it just for reading, or for both reading and writing. The time spent by users for writing articles and the frequency of usage was also considered. The other 50% pertain to the kind of content that they like to read, for which the following options were given: random things, people's experiences in social situations, problems in life, the life that the user wants to lead. This section will give a deep insight into a user's personal space and their general level of happiness to the machine learning model.

In this work, user demographics is not included in the feature set. This is primarily because the study presents the impact of social media usage on the mental health of any individual who belongs to the focused age group. However, the impact of considering gender information on the results can be studied as a part of future work on the subject.

3.6 Data Set Building & Measuring depressive behaviour

To allow the supervised learning algorithms for training the machine learning models, the dataset has been augmented with a ground truth label corresponding to each entry. This label indicates whether or not there is a possibility of the participant being depressed. All participants were thus divided into two sets, one representing individuals who were possibly depressed, second represented those individuals who were most likely not depressed.From among the total 300 participants, 40% (120 participants) were indicative of possibly being depressed and 60% (180 participants) were indicative of lack of the possibility of suffering from depression.

In this work, GDS depression scale [13] as a means to assess the user's mental health condition is used. Owing to the preliminary nature of the study, this test suffices in classifying data points into de-pressed or not depressed classes.

Table 1. Summary of items in Online Survey

Item	Description
Demographic Details	Including *age* and *gender* of the participants.
Facebook and Instagram usage preferences	Questions pertaining to the participant's usage habits with respect to the most popular social networking websites Facebook and Instagram.
Involvement in tracking influencer and celebrity lives	Questions regarding the participant's interest in celebrity lives and their motivation to follow them across different media outlets.
Hours spent streaming video and the type of content streamed	Questions aimed at collecting information on genres a participant likes for entertainment purposes.
Digital gaming involvement and thoughts on video game violence	Questions that gauge the participant's interest in various popular online games and their views on video game violence.

Time spent reading and engaging in discussions on Quroa, Reddit under several categories	The participant's interests and habits on the most popular online forums Quora and Reddit.
GDS Short Form	A Stanford form that is used to assess the participant using a series of 15 questions about their state of mind with the help of a mood scale.

The form assesses the participant using a series of 15 questions about their state of mind with the help of a mood scale. The participants can answer all these questions with either a yes or a no where one of the answers indicates depression and adds a point to the final score. According to the scoring guide-lines of the form a score of more than 5 indicates the presence of depression while a score of greater than 10 strongly suggests that the participant is depressed and should follow up with a doctor. For this study it has been assumed a score of greater than 5 to be indicative of depression. Table 1 summarizes the questions in the online survey.

4. PROPOSED TECHNIQUE

This stage constructs prediction model for depression recognition by considering social media activities of users as input features. Considering the input features, each user is labeled a binary output class (1 indicating possible presence of depression and 0 meaning no such result was found). The task of the classifier is to predict the corresponding label for each user. In this work, seven popular classifiers are used: Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM), k-Nearest Neighbor (KNN), Random Forest (RF), Decision Tree (DT) and Multi-layer Perceptron (MLP). 70% of the instances from the dataset were used for training while the remaining 30% were used for testing.

Following are the steps to the procedure for training the classification models:

Step 1: Initialize the following variables:

(i) accuracy_list: Stores test accuracy produced by each classifier.

(ii)prediction_probability_list: Stores prediction probability of positive out-come for every classification algorithm.

Step 2: Split data into X_train, y_train, X_test and y_test.

Step 3: Scale X_train and X_test.

Step 4: Repeat steps 5-8 for each of the above-mentioned classification algorithms

Step 5: Perform training by the using the current classification algorithm.

Step 6: Utilize the trained classifier to calculate prediction probability providing X_test as input.

Step 7: Calculate test accuracy using the current classifier and append it to accuracy_list.

Step 8: Add prediction probability of the positive outcome for the current algorithm to prediction_probability_list.

Step 9: For each classification algorithm, evaluate AUC and plot the ROC curve.

Step 10: Print test accuracies for each classification algorithm from accuracy_list.

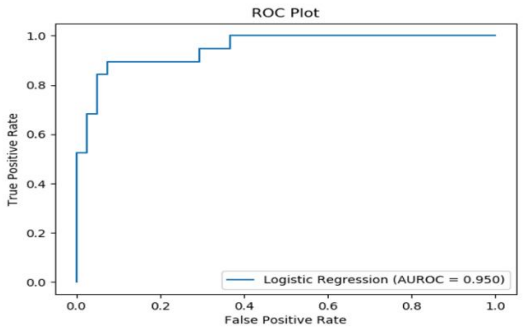

Figure 1: ROC curve for Logistic Regression

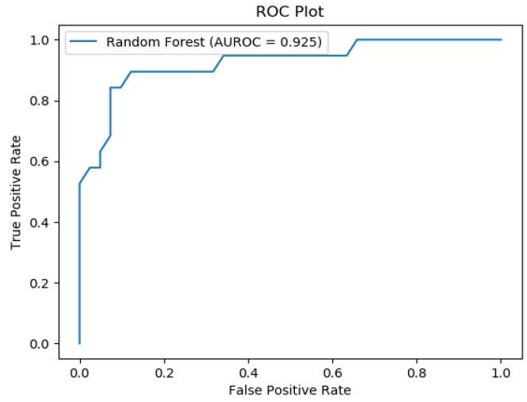

Figure 2: ROC curve for Random Forest

5. RESULT & DISCUSSION

The results obtained by applying different classification algorithms have been listed in Table 2. The features extracted from the participants' online habits and their depression score were used as in-put for the algorithms.

The obtained results show that the proposed method was able to detect the presence of depression in test data with the maximum accuracy of 90% using logistic regression and 88.3% using support vector machine.

ROC (Receiver Operating Characteristics) curves and their associated AUC (Area Under the Curve) are also used for the purpose of evaluation. A high value of the AUROC metric strongly indicates higher accuracy for the model.

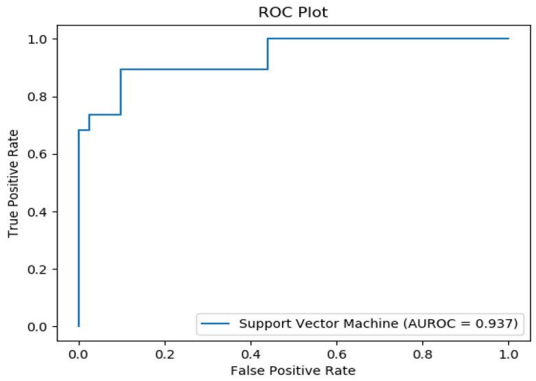

Figure 3: ROC curve for SVM

Table 2. Performance of the Binary Classifiers

Model	Accuracy	AUROC
Logistic Regression	90%	0.950
Naive Bayes	76.6%	0.726
Support vector machine	88.3%	0.937
K nearest neighbor	86.6%	0.917
Random Forrest	85%	0.925
Decision Tree	81.6%	0.817
Multi-layer perceptron	88.3%	0.920

Maximum AUC of 0.95 was yielded by Logistic Regression, followed by 0.937 for Support Vector Machine. While AU-ROC values of 0.7 and higher are considered strong effects in applied psychology and prediction of future behavior, these values make the model fit for employment even in medical diagnosis, where very high AUROCs (>=0.95) are sought [16]. ROC curves of the best performing classifiers are shown in Figure 1, Figure 2 and Figure 3. These report high values of AUROC (>0.9), which indicate "Outstanding discrimination" [17].

To study the correlation between the various input features and the output class, the chi-square test of independence with an alpha value of 0.05 for the said features has been done. It has been observed that there is a strong dependence between the output and input features. The features with the lowest p-values were then used for calculating Pearson correlation coefficients.The obtained results are highlighting towards a positive correlation between various types of social media activities and depressive tendencies. Particularly it has been found that p-values for "Hours spent streaming video", "Time spent playing digital games" (particularly violence-based games), and "Elevated interest in Instagram followers" were the features which had the strongest positive correlation. Interestingly, it has also been inferred that there is an inverse correlation between "Time spent watching news related content" and depressive tendencies which conform to the general consensus that people suffering from depression tend to lose interest in the world around them.

6. CONCLUSION & FUTURE SCOPE

This work proposes a machine learning based approach to prediction of general depression amongst social media users. Data about social media habits of participants has been used to extract 30 features, which act as input to the proposed methods. Results obtained predict the presence of general depression with a reasonable accuracy. A strong inverse correlation between unrestricted social media usage and mental health of the user has also been found, which helped to find insights into the effects of social media on mental health.

The current scope of this study has been limited to users in developing nations, although this approach could be extended to the global population. A strong positive correlation between the usage of social media and general deterioration of mental health has been calculated. These findings suggest that an automated system may be developed and deployed with a trained model that can detect the user's deteriorating mental health and automatically suggest changes in social media habits for a healthier life.Due to the preliminary nature of this study and the sensitive nature of the data, the collected dataset was relatively local. The same study might be conducted with a higher number of participants to make the findings more holistic. Future researchers should also try to increase the number and precision of the features used since these are subject to change with the locale of the population under consideration

REFERENCES

1. Krishnan, V. (2019, August 21). **How much time do Indians spend on social media?** Retrieved July 07, 2020, from https://www.thehindu.com/news/national/how-much-time-do-indians-spend-on-social-media/article29201363.ece

2.Hunt, M. G., Marx, R., Lipson, C. & Young, J. (2018). **No More FOMO: Limiting Social Media Decreases Loneliness and Depression**. Journal of Social and Clinical Psychology, 37, 751--768. doi: 10.1521/jscp.2018.37.10.751

3. Tiwari, R., Negandhi, H., &Zodpey, S. (2019). **Forecasting the future need and gaps in requirements for public health professionals in India up to 2026**. WHO South-East Asia Journal of Public Health,8(1), 56. doi:10.4103/2224-3151.255351

4. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). . In International AAAI Conference on Web and SocialMedia. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6124

5. Woods, H. C., & Scott, H. (2016). **#Sleepyteens: Social media use in adolescence is associated with**

poor sleep quality, anxiety, depression and low self-esteem. Journal of Adolescence,51, 41-49. https://doi.org/10.1016/j.adolescence.2016.05.008

6. Rost, K., Zhang, M., Fortney, J., Smith, J., Coyne, J., & Richard Smith, G. (1998). **Persistently poor outcomes of undetected major depression in primary care**. General Hospital Psychiatry, 20(1), 12–20. doi:10.1016/s0163-8343(97)00095-9

7.**Depression**. (n.d.). Retrieved July 07, 2020, from https://www.who.int/news-room/fact-sheets/detail/depression

8. Islam, M. R., Kabir, M. A., Ahmed, A., Kamal, A. R., Wang, H., &Ulhaq, A. (2018). **Depression detection from social network data using machine learning techniques.** Health Information Science and Systems,6(1). doi:10.1007/s13755-018-0046-0

9. Aggarwal, S., Saluja, S., Gambhir, V., Gupta, S., &Satia, S. P. S. (2019). **Predicting likelihood of psycho-logical disorders in PlayerUnknown's Battlegrounds (PUBG) players from Asian countries using Supervised Machine Learning**. Addictive Behaviors, 106132. doi:10.1016/j.addbeh.2019.106132

10. Braithwaite, S. R., Giraud-Carrier, C., West, J., Barnes, M. D., & Hanson, C. L. (2016). **Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality**. JMIR Mental Health, 3(2), e21. https://doi.org/10.2196/mental.4822

11. Wongkoblap, A., Vadillo, M. A., &Curcin, V. (2017). **Researching Mental Health Disorders in the Era of Social Media: Systematic Review**. Journal of Medical Internet Research, 19(6), e228. https://doi.org/10.2196/jmir.7215

12. Morrison, C. M., & Gore, H. (2010). The **Relationship between Excessive Internet Use and Depression: A QuestionnaireBased Study of 1,319 Young People and Adults**. Psychopathology, 43(2), 121–126. https://doi.org/10.1159/000277001

13. **Geriatric Depression Scale Short Form English Scoring**. (n.d.). Retrieved July 07, 2020, from https://web.stanford.edu/~yesavage/GDS.english.short.score.html

14. **Digital Trends 2019 & Social Media Landscape in India.** (2020, June 29). Re-trieved July 07, 2020, from https://sannams4.com/digital-and-social-media-landscape-in-india/

15. **List of countries by number of Internet users.** (2020, July 04). Retrieved Ju-ly 07, 2020, from https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

16. Rice, M. E., & Harris, G. T. (2005). **Comparing effect sizes in follow-up studies: ROC Area**, Cohen's d, and r. Law and Human Behavior, 29(5), 615-620. doi:10.1007/s10979-005-6832-7

17. Lemeshow, S., Sturdivant, R. X., & Hosmer, D. W. (2013). **Applied Logistic Regression** (Wiley Series in Probability and Statistics). p. 177. Wiley.