ISSN: 2473-4901
v12 n6709

Predicting Mental Health Risks of Social Media Use with Machine Learning

Jie Du
dujie@gvsu.edu
Grand Valley State University
Allendale, MI 49401

Hayden Wimmer
hwimmer@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Vishakha Nikam
vn01581@georgiasouthern.edu
Georgia Southern University
Atlanta, GA 30302

Abstract

Social media has become deeply embedded in everyday life, raising concerns about the way it interferes with mental health. This paper investigates the prediction of mental-health outcomes, including depression, distraction, and sleep issues, based on social media usage patterns using machine-learning approaches on an open-source Kaggle dataset containing self-reported demographics, usage patterns, and mental-health indicators. Four machine-learning approaches, including Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and a Multi-Output Classifier, were applied to predict multiple mental-health outcomes including depression, distraction, and sleep issues simultaneously. The evaluation used accuracy, precision, recall, F1-score, weighted F1-score, and confusion matrices. LR, SVM, and XGBoost achieved modest performance (around 30–56% accuracy), and struggled particularly with multi-class severity levels. In contrast, the Multi-Output Classifier achieved an accuracy of 82% for depression, 87% for distraction, and 68% for sleep issues. The results suggest that mental-health outcomes related to social media use are multidimensional, making it more effective to analyze all the interrelated symptoms together rather than treating them individually.

Keywords: Social Media, Mental Health, Machine Learning, Data Analysis

 ISSN: 2473-4901
v12 n6709

Predicting Mental Health Risks of Social Media Use with Machine Learning

Jie Du. Hayden Wimmer and Vishakha Nikam

1. INTRODUCTION

Social media platforms such as Facebook, X, Snapchat, Instagram, YouTube, Discord, Pinterest, and TikTok are widely used for communication, sharing, and searching for information. These platforms are shaping modern lifestyles as their use continues to grow, meanwhile they have resulted in negative impacts on mental health, such as anxiety, depression, distraction, and excessive time spent online (Vidal et al., 2020). Its excessive use among adults, especially young adults, affects relationships, weakens bonding between family members and friends, and reduces physical and social interactions (Shannon et al., 2022). Among parents, there are increased concerns that their children's use of social media may interfere with their sleep and academic responsibilities and expose them to risks such as cyberbullying and inappropriate interactions (Hernandez et al., 2024).

Researching social media's impact is essential, as it shapes moods and online behavior, and understanding this link can guide healthier digital habits and greater mental-health awareness. However, as the volume of data on online behaviors is increasing at a rapid pace, conventional research approaches find it difficult to reveal the intricate connections between social media use and mental health. Traditional studies primarily rely on surveys, questionnaires, and statistical analyses to examine the relationship between social media use and mental health. While effective for small-scale analyses, these approaches are limited in capturing complex, multidimensional patterns and interactions within large behavioral datasets. Machine Learning (ML) is a current, potent method of examining immense, multidimensional information and revealing more intricate associations that cannot be readily observed by hand. With the help of ML, researchers can recognize behavioral patterns, risk levels, and forecast the influence of specific usage patterns on emotional well-being (Liu et al., 2022 and Ahmed et al., 2022).

This study aims to evaluate the effectiveness of ML approaches for predicting mental-health outcomes based on social media usage patterns. By comparing the performance of multiple ML models, this study identifies the most effective approach for predicting mental health outcomes associated with social media use. Predictive models including Logistic Regression (LR), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Multi-Output Classifier with Python libraries are used to predict mental issues related to the use of social media. A comparative analysis is conducted to identify the most effective model to make trustworthy predictions on behaviors. The findings highlight the need for awareness and digital literacy to promote safe and balanced social media engagement. This study helps establish the groundwork for future work on how to forecast psychological and behavioral trends.

The remainder of this paper is structured as follows: Section 2 reviews the relevant literature. Section 3 details the methodology, including dataset preparation, model development, and evaluation metrics. Section 4 presents the experimental results, followed by a discussion in Section 5 that highlights implications, limitations, and directions for future research. Section 6 concludes the study.

2. LITERATURE REVIEW

2.1 Evolution of Social Media

In the last 20 years, social media has evolved as a potent digital space that influences the perceptions of individuals and others. Aicher et al. (2021) conducted a structured literature review of social media research from 1994 to 2019 and found that early definitions emphasized connectivity among people and virtual communities, whereas later definitions increasingly focus on user-generated content and multifunctional platforms. Social media offers forums of creativity, education, and social awareness. It is used by many creators to spread positive messages about mental health, knowledge, and to connect with the various communities (Vilarinho-Pereira et al., 2021). However, the border between normal use

ISSN: 2473-4901
v12 n6709

and detrimental overexposure is thin. Social media allows people to connect with others all over the world, while equally increasing insecurity and emotional exhaustion among them (Sheng et al., 2023).

2.2 Online Behaviors

On the social media era, natural boundaries have been eliminated. People can easily comment, write, or criticize others behind the screen, often without considering the emotional impact of their words. It has led to the emergence of behaviors like cyberbullying and online trolling, representing forms of digital harassment characterized by repeated aggressive behaviors that can cause significant psychological harm to victims (Aydın et al., 2021). Cyberbullying has been consistently linked to negative mental outcomes including increased anxiety, depression, stress, self-harm, and suicidal ideation among adolescents (Hu et al., 2021 and Bottino et al., 2015). In addition to emotional coverage, the prolonged use of social media is a way of causing behavioral problems (Sheng et al., 2023). The effects of such influences are especially dangerous to adolescents and young adults, who are still developing their identity and social confidence, and tend to grow more anxious, lonely, and less self-accepting (Vidal et al., 2020 and Shannon et al., 2022).

2.3 Social Media and Mental Health

Researchers have extensively investigated the relationship between social media use and mental health outcomes. Kuss and Griffiths (2011) excavated 43 research topics regarding social networking service (SNS) addiction and found that people primarily use these platforms to stay connected with their real-life friends. They caution that the SNS addiction may harm school achievements, initiate relationship fights, and decrease real-life socialization. Coyne et al. (2020) followed approximately 500 teenagers, aged 13 to 20, over eight years to determine whether the use of social media technology influenced the development of depression and anxiety. Their findings suggest that greater overall use was associated with poor mental health. Naslund et al. (2020) presented an objective perspective of the advantages and disadvantages of social media on mental health. Bashir and Bhat (2017) discussed how binge use of social media can significantly increase anxiety, stress, and depression, and loneliness due to social comparison, online warring, while reducing face-to-face conversations. Garg (2023) found that moderate use enhances a sense of connectedness, whereas heavy use is associated with increased anxiety and depression. Karim et al. (2020) reviewed prior studies to investigate the relationship between social media and anxiety, depression, and self-esteem. O'Reilly et al. (2019) investigated whether a social media-based intervention could serve as a cost-effective method for promoting adolescents' mental health. Gao et al. (2020) suggested that regulated, school-based interventions using social media could positively influence the mental health of teenagers.

2.4 ML Approaches

With the rapid growth of online behavioral data, ML offers a powerful approach to uncover complex, multidimensional associations between social media use and mental health that traditional methods struggle to detect. A systematic review (Liu et al., 2022) summarizes the findings of previous studies published between January 1990 to December 2020 that applied ML methods to social media text data to detect depression. Their findings demonstrate that ML-based analysis of social media text data is both feasible and effective for detecting mental health conditions. Ahmad et al. (2022) examined 54 studies that applied ML techniques to social media data to detect signs of anxiety and depression. Kim et al. (2020) developed a deep learning model to identify a user's mental state based on his/her posting information on Reddit. Anand et al. (2024) investigated the influence of social media use on students and teenagers using survey analysis, ANOVA, correlation statistics, and ML algorithms, including SVM and LR. Their findings showed that social media usage patterns are associated with mental-health outcomes, including anxiety, depression, and self-esteem, demonstrating the potential of ML models for analyzing these relationships. Lamba et al. (2026) investigated explainable machine learning approaches for mental health prediction using social media behavior data. Their study evaluated models including Logistic Regression, SVM, and Random Forest with SHAP and LIME techniques to improve model interpretability.

Although previous studies have explored the relationship between social media use and mental health using statistical methods and ML approaches, several limitations remain. Many existing studies focus on

ISSN: 2473-4901
v12 n6709

a single mental-health outcome or analyze text-based social media data, while fewer studies examine multiple behavioral indicators such as depression, distraction, and sleep issues together. Additionally, there is limited comparison of different ML models on the same dataset to determine their effectiveness in predicting multiple mental-health outcomes. Therefore, this study addresses these gaps by applying and comparing multiple ML models, including LR, SVM, XGBoost, and Multi-Output Classifier, to predict mental-health outcomes associated with social media usage patterns.

3. METHODOLOGY

The dataset "Social Media and Mental Health" used in this study was collected by Souvik Ahmed and Muhesena Nasiha Syeda and downloaded from the Kaggle platform (https://www.kaggle.com/datasets/souvikahmed071/social-media-and-mental-health). The dataset was published and updated in 2023 and contains survey responses collected from 481 individuals in Dhaka, Bangladesh. The dataset has also been used in previous research to investigate the association between long-term social media use and mental health outcomes (Tao, 2024). This dataset consists of 481 survey-based responses containing individuals social media usage patterns and various mental-health indicators. The dataset contains demographics, social media usage patterns, and self-reported mental health indicators. The dataset was cleaned to ensure consistency before analysis. After cleaning, variables were sorted into numerical and categorical sets.

3.1 Model Development

This study utilizes four ML models to predict mental health outcomes based on the use of social media, including LR, SVM, XGBoost, and a Multi-Output Classifier. Each model was enclosed in scikit-learn Pipelines, which contain preprocessing steps alongside the learning algorithm. This will remove data leaks and ensure that transformations occur sequentially in the correct order.

3.1.1 Logistic Regression

The reason why LR was selected as the baseline algorithm is that it is one of the easiest models to describe. It is widely used in mental-health and social-science research, and it allows us to visualize clearly the movements of each predictor, towards or away, the odds of a negative psychological outcome. In contrast to linear regression, which provides a continuous value, LR provides the likelihood of a person falling into one of certain categories, such as experiencing depressive symptoms or describing oneself as highly distracted.

This model uses the input features of hours of screen time, comparison behavior, online validation seeking, or sleep-related habits to create a number between 0 and 1 through the sigmoid function:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)}}$$

Here,

- $Y = 1$ represents the presence of the mental-health condition being predicted,
- $X_1 ... X_n$ are the features, and
- $\beta_0 ... \beta_n$ are model coefficients learned from the data.

All the coefficients prove the fact that a feature increases or decreases the chances of having a certain effect. It is popular in psychology adjacent studies because this interpretability is what makes it so desirable; the reason why it matters in the outcome of a prediction is as important as the accuracy itself.

3.1.2 SVM

SVM was adopted because it is effective in dealing with psychological and behavioral data, and where the lines separating the classes may not be crisp or straight. A combination of habits, emotional states, and online behavior influences social media effects on mental health, and thus, relationships can become rather complicated. In SVM, data were grouped into two categories by determining the optimal boundary, which is a hyperplane, that will maximize the margin between the two groups. The distance

ISSN: 2473-4901
v12 n6709

between the hyperplane and the nearest data points, which are referred to as support vectors, was just the margin.

The decision function for SVM is:

$$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b)$$

where

- $K(x_i, x)$ is the kernel function,
- y_i are class labels,
- α_i are learned weights,
- b is the bias term.

The SVM was trained to make a classification, such as the level of distraction and depressive symptoms. Hyperparameters were adjusted by means of C (strength of regularization) and gamma (width of the kernel) to achieve a compromise between generalization and accuracy. The model was able to capture the patterns that were not possible under LR, particularly where mental-health outcomes were nonlinear in their interactions.

3.1.3 XGBoost Classifier

XGBoost is one of the most powerful ML algorithms to use in tabular social-behavioral data, particularly when the relationships between the features are subtle or hierarchical. XGBoost creates a set of decision trees, with the successive trees trying to rectify the mistakes of its predecessors. The sequential boosting structure enables the model to discover complex associations among social-media activity and the impact on mental health. It finds nonlinear relationships and interactions between features without requiring manual engineering.

$$\mathcal{L} = \sum_{i} l(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k)$$

The regularization term:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \, || \, w \, ||^2$$

The model was conditioned to forecast such results as depressive symptoms and sleep disruption. Learning rate, max depth, and number of trees were parameters that were tuned according to their performance.

3.1.4 Multi-Output Classifier

A multi-output classifier was used to capture the fact that mental-health impacts tend to be co-occurring. A common classifier predicts a single outcome, whereas a multi-output classifier predicts multiple labels related to mental health simultaneously. A multi-output classifier uses a different model per label, such as distraction, depression, and sleep problems, yet has the same underlying inputs. This arrangement captures the co-occurrence of symptoms that is consistent with psychological studies that indicate mental-health indicators are hardly expressed in isolation (Brown et al., 2001 and Starr et al., 2012). If using LR internally, the model optimizes:

$$P(Y_1, Y_2, \dots, Y_k \mid X) = \prod_{i=1}^{k} P(Y_i \mid X)$$

where each Y_i represents a different mental-health outcome.

Although each classifier is trained independently, they share the same input features, allowing the model to detect patterns across variables.

ISSN: 2473-4901
v12 n6709

3.2 Model Evaluation

All the models were compared using accuracy, precision, recall, the F1-score, and the weighted F1-score. Accuracy provides an overall measure of correct predictions, while precision and recall assess the balance between identifying true symptom cases and avoiding false alarms.

Accuracy measures the overall proportion of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

Precision evaluates how many of the positive predictions are actually correct. This matters when false alarms (incorrectly predicting symptoms) need to be minimized.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also called Sensitivity) measures how well the model identifies actual cases of symptoms.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1-score combines precision and recall into one balanced metric by using their harmonic mean:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Because symptom classes may be imbalanced, the weighted F1-score is used to better reflect performance across all classes. It averages the F1-scores of each class while weighing them by their support (the number of true samples in each class).

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{k} w_i \times \text{F1}_i$$

where

$w_i = \dfrac{\text{Support of class } i}{\text{Total samples}}$

k= number of classes

4. RESULTS

Each machine learning classifier was trained using the selected input features to learn the relationship between the predictor variables and the target classes. During testing, the trained model predicted one of the five severity classes (Class 1–Class 5) for each sample based on the learned patterns. These predictions were then compared with the actual labels to evaluate model performance.

The target variables (depression, distraction, and sleep issues) were represented as five classification categories based on the original survey responses. The classes range from Class 1 to Class 5, where each class corresponds to the severity level reported by participants. The confusion matrices present the actual classes in rows and predicted classes in columns, with diagonal values representing correctly classified observations and off-diagonal values representing misclassifications. A stronger concentration of values along the diagonal indicates better classification performance.

The performance of each classifier was evaluated using confusion matrices and classification metrics. Precision measures the proportion of correctly predicted positive cases, recall measures the ability of the model to identify actual positive cases, and the F1-score provides a balance between precision and

ISSN: 2473-4901
v12 n6709

recall. Support indicates the number of samples belonging to each class. Accuracy, macro average, and weighted average are reported to provide an overall evaluation of model performance.

4.1 Logistic Regression Model

The confusion matrix (see Table 1) shows that the LR model performs best on Class 1 and Class 5, while exhibiting substantial misclassification among mid-level classes. This indicates that it is difficult to distinguish similar behavioral states. Table 2 shows the performance of the LR model on classifying the five levels of depression. The LR model achieves approximately 38% accuracy; however, precision and recall vary significantly across different classes. Class 3, which has 24 support instances, shows the lowest performance consistent across precision, recall, and F1-score. Performance on Class 1 and Class 5 is slightly better, but the improvement is not consistent. Class 5, with 28 support instances, has a low recall of 0.36. suggesting many instances in this category has been misclassified into adjacent categories. Overall, the results suggest LR has limited capability to reliably distinguish similar levels of depression in this dataset.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	9	1	2	2	0
Class 2	2	7	3	2	0
Class 3	2	4	4	9	5
Class 4	0	2	6	7	2
Class 5	0	2	7	9	10

Table 1. Logistic Regression Confusion Matrix – Depression

Class	Precision	Recall	F1-Score	Support
1	0.69	0.64	0.67	14
2	0.44	0.50	0.47	14
3	0.18	0.17	0.17	24
4	0.24	0.41	0.30	17
5	0.59	0.36	0.44	28
Accuracy			0.38	97
Macro Avg	0.43	0.42	0.41	97
Weighted Avg	0.42	0.38	0.39	97

Table 2. Logistic Regression Performance Metrics – Depression

4.1.1 Evaluating Model for Target: Distraction

The confusion matrix (see Table 3) shows strong performance on Class 3 and Class 5 and substantial misclassification on Class 2 and Class 4. It indicates that the model correctly identifies the dominant class reasonably well but still exhibits noticeable confusion across neighboring categories, suggesting limited precision. Table 4 shows the detailed performance measures. The LR model achieves approximately 49% accuracy in predicting the levels of distraction. The model performs reasonably well for Classes 1, 3, and 5 (the higher and mid-high categories), whereas performance for Classes 2 and 4 is weak. This indicates that the model struggles to distinguish small differences in distraction. The results suggest that the LR model is only moderately dependable for predicting distraction patterns.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	4	2	0	0	0
Class 2	0	5	6	2	0
Class 3	1	6	22	6	4
Class 4	0	0	8	6	5
Class 5	0	0	3	6	11

Table 3. Logistic Regression Confusion Matrix – Distraction

Class	Precision	Recall	F1-Score	Support
1	0.80	0.67	0.73	6
2	0.38	0.38	0.38	13
3	0.56	0.56	0.56	39

 ISSN: 2473-4901
v12 n6709

Class	Precision	Recall	F1-Score	Support
4	0.30	0.32	0.31	19
5	0.55	0.55	0.55	20
Accuracy			0.49	97
Macro Avg	0.52	0.50	0.51	97
Weighted Avg	0.50	0.49	0.50	97

Table 4. Logistic Regression Performance Metrics – Distraction

4.1.2 Evaluating Model for Target: Sleep Issues

The confusion matrix (see Table 5) shows extensive overlap between categories, indicating that the model struggles significantly to capture sleep-related patterns. Table 6 shows the performance of the LR model on classifying the levels of sleep issues. This model performs poorly on predicting sleep-related issues, with a 31% accuracy. It exhibits low precision and recall for Classes 1, 2, 4, and 5. Notably it fails to correctly classify any instances of Class 3. These results suggest that LR is not a suitable tool for predicting sleep-related patterns.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	7	3	1	2	4
Class 2	2	7	1	5	7
Class 3	1	2	0	5	6
Class 4	2	2	3	2	9
Class 5	3	5	0	4	14

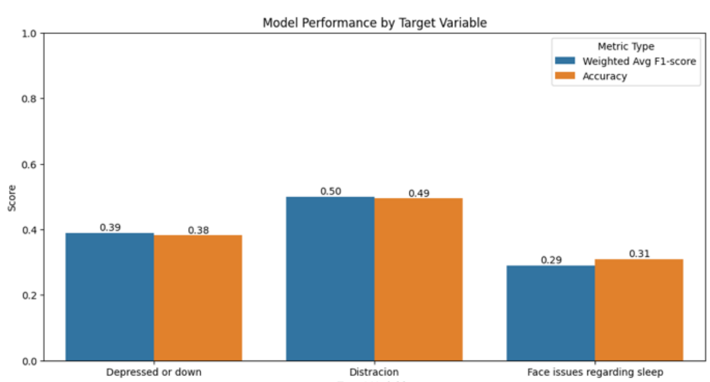
Table 5. Logistic Regression Confusion Matrix – Sleep Issues

Class	Precision	Recall	F1-Score	Support
1	0.47	0.41	0.44	17
2	0.37	0.32	0.34	22
3	0.00	0.00	0.00	14
4	0.11	0.11	0.11	18
5	0.35	0.54	0.42	26
Accuracy			0.31	97
Macro Avg	0.26	0.28	0.26	97
Weighted Avg	0.28	0.31	0.29	97

Table 6. Logistic Regression Performance Metrics – Sleep Issues

4.1.3 Comparison Across Outcomes

Figure 1 illustrates the performance of the LR model across three mental health outcomes. The LR model demonstrates moderate predictive accuracy in predicting detrimental effects of social media on mental health. In particular, the accuracy of the model is 38% for depression classification, 49% for distraction classification, and 31% for sleep issues classification. These findings suggest that social media may contribute to mental health issues, although other psychological and environmental factors are also likely involved.



 ISSN: 2473-4901
v12 n6709

Figure 1. Model Accuracy and F1-Score by Target Variable (Logistic Regression)

4.2 SVM Model

4.2.1 Evaluating Model for Target: Depression

For the *Depressed or Down* target, the confusion matrix (see Table 7) shows slightly improved correct classifications compared to LR, but misclassification across similar classes remains high, indicating that the model is unable to distinguish similar depression states. When it comes to detecting the degree of depression, the SVM model only achieved approximately 38% (see Table 8). The model performs reasonably well for Classes 1 and 2, but its performance deteriorates substantially by Class 3. Classes 1 and 5 show somewhat better precision, but their recall remains inconsistent.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	8	2	3	1	0
Class 2	2	7	3	2	0
Class 3	2	5	6	5	6
Class 4	0	3	3	6	5
Class 5	1	1	7	9	10

Table 7. SVM Confusion Matrix – Depression

Class	Precision	Recall	F1-Score	Support
1	0.62	0.57	0.59	14
2	0.39	0.50	0.44	14
3	0.27	0.25	0.26	24
4	0.26	0.35	0.30	17
5	0.48	0.36	0.41	28
Accuracy			0.38	97
Macro Avg	0.40	0.41	0.40	97
Weighted Avg	0.40	0.38	0.38	97

Table 8. SVM Performance Metrics – Depression

4.2.2 Evaluating Model for Target: Distraction

For the *Distraction* target, the model demonstrates stronger class recognition with a more pronounced diagonal, although confusion with adjacent classes persists (see Table 9). The SVM achieves an accuracy of 56%, showing an improvement over LR, as shown in Table 10. It performs best for Class 5 and shows strong performance for Class 3, while the performance on Class 2 remains weak. Generally, the SVM performs well in terms of distraction, and the separation between classes is improved.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	4	1	1	0	0
Class 2	2	4	6	1	0
Class 3	2	4	23	7	3
Class 4	0	0	9	9	1
Class 5	1	0	2	3	14

Table 9. SVM Confusion Matrix – Distraction

Class	Precision	Recall	F1-Score	Support
1	0.44	0.67	0.53	6
2	0.44	0.31	0.36	13
3	0.56	0.59	0.57	39
4	0.45	0.47	0.46	19
5	0.78	0.70	0.74	20
Accuracy			0.56	97
Macro Avg	0.54	0.55	0.53	97
Weighted Avg	0.56	0.56	0.56	97

ISSN: 2473-4901
v12 n6709

Table 10. SVM Performance Metrics – Distraction

4.2.3 Evaluating Model for Target: Sleep Issues

In the case of sleep issues, the confusion matrix (see Table 11) shows better performance on extreme classes (Class 1 and Class 5) but misclassifications on Class 3. It reveals highly scattered predictions, indicating that the model fails to effectively learn distinguishing features for this category. The SVM achieves an accuracy of 31%, indicating a poor predictive performance, as shown in Table 12. The performance on middle classes, especially Class 3 is extremely poor. The performance on Class 5 only appears a bit more optimistic, yet very weak.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	8	1	1	3	4
Class 2	2	7	3	6	4
Class 3	3	3	1	2	5
Class 4	3	3	1	2	9
Class 5	2	4	1	7	12

Table 11. SVM Confusion Matrix – Sleep Issues

Class	Precision	Recall	F1-Score	Support
1	0.44	0.47	0.46	17
2	0.39	0.32	0.35	22
3	0.14	0.07	0.10	14
4	0.10	0.11	0.11	18
5	0.35	0.46	0.40	26
Accuracy			0.31	97
Macro Avg	0.29	0.29	0.28	97
Weighted Avg	0.30	0.31	0.30	97

Table 12. SVM Model Performance Metrics – Sleep Issues

4.2.4 Comparison Across Outcomes

The SVM model presented a reasonable amount of mixed output (see Figure 2). The validity was approximately 38% with depressed or down feelings, approximately 56% with distraction, and 31% with sleep-related problems. The model achieves its highest performance on distraction, while performance on depression and sleep is comparatively poor. Therefore, it appears that social media is more likely to be associated with distraction, whereas the effects it has on depression or sleep-related issues are likely influenced by other mental issues and circumstances.

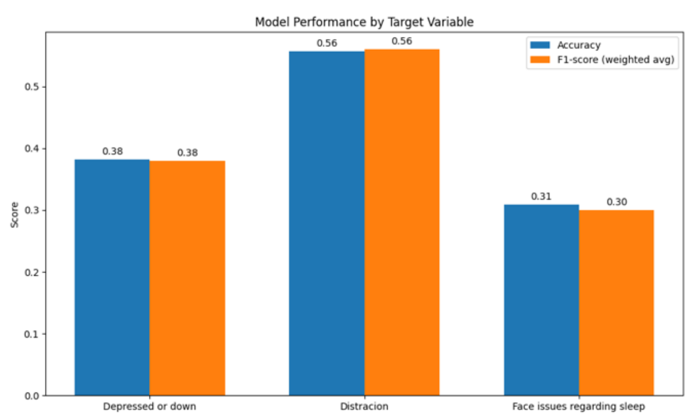


Figure 2. Model Accuracy and F1-Score by Target Variable (SVM)

ISSN: 2473-4901
v12 n6709

4.3 XGBoost Model

4.3.1 Evaluating Model for Target: Depression

For the *Depressed or Down* target, the confusion matrix (see Table 13) shows moderate correct predictions but significant off-diagonal values, suggesting difficulty in modeling class boundaries. As shown in Table 14, XGBoost achieves an overall accuracy of 37% when predicting depressive feelings. The lower classes, such as Class 1, perform better; however, the performance on the intermediate classes, in particular, Class 3, is not good. The results suggest poor this target can be relied on only to a small extent in the model.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	9	1	2	2	0
Class 2	2	7	4	1	0
Class 3	1	5	4	7	7
Class 4	0	2	4	6	5
Class 5	1	1	8	8	10

Table 13. XGBoost Model Confusion Matrix – Depression

Class	Precision	Recall	F1-Score	Support
1	0.69	0.64	0.67	14
2	0.44	0.50	0.47	14
3	0.18	0.17	0.17	24
4	0.25	0.35	0.29	17
5	0.45	0.36	0.40	28
Accuracy			0.37	97
Macro Avg	0.40	0.40	0.40	97
Weighted Avg	0.38	0.37	0.37	97

Table 14. XGBoost Model Performance Metrics – Depression

4.3.2 Evaluating Model for Target: Distraction

For the *Distraction* target, the confusion matrix (shown in Table 15) reflects weak class separation with predictions dispersed across multiple classes, indicating poor learning performance. The XGBoost model shows weak performance for the distraction target, with an accuracy of only 35% (see Table 16). Most of the classes show scattered performance, with precision and recall ranging from low to moderate. Among them, Class 5 demonstrates the strongest performance. Generally, there is quite an inconsistent performance between the classes, and thus, XGBoost does not add much predictive value to this target.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	3	5	1	1	0
Class 2	3	7	2	4	1
Class 3	1	4	5	9	4
Class 4	2	4	6	8	7
Class 5	0	0	1	8	11

Table 15. XGBoost Model Confusion Matrix – Distraction

Class	Precision	Recall	F1-Score	Support
1	0.33	0.30	0.32	10
2	0.35	0.41	0.38	17
3	0.33	0.22	0.26	23
4	0.27	0.30	0.28	27
5	0.48	0.55	0.51	20
Accuracy			0.35	97
Macro Avg	0.35	0.36	0.35	97
Weighted Avg	0.35	0.35	0.34	97

Table 16. XGBoost Model Performance Metrics – Distraction

ISSN: 2473-4901
v12 n6709

4.3.3 Evaluating Model for Target: Sleep Issues

For the *Sleep Issues* target, the model shows slight improvement in correct classifications compared to other baseline models, but overall misclassification remains high (see Table 17). The XGBoost model achieves an overall accuracy of 35% for sleep-related problems, as shown in Table 18. While Classes 1 and 5 perform slightly better, Class 3 shows a very low recall. Accuracy is poor in general, and therefore, XGBoost is extremely ineffective in this prediction problem.

True \ Pred	*Class 1*	*Class 2*	*Class 3*	*Class 4*	*Class 5*
Class 1	8	1	2	2	4
Class 2	2	7	1	6	6
Class 3	1	3	1	5	4
Class 4	2	2	2	3	9
Class 5	3	4	2	2	15

Table 17. XGBoost Model Confusion Matrix – Sleep Issues

Class	Precision	Recall	F1-Score	Support
1	0.50	0.47	0.48	17
2	0.41	0.32	0.36	22
3	0.12	0.07	0.09	14
4	0.17	0.17	0.17	18
5	0.39	0.58	0.47	26
Accuracy			0.35	97
Macro Avg	0.32	0.32	0.31	97
Weighted Avg	0.34	0.35	0.34	97

Table 18. XGBoost Model Performance Metrics – Sleep Issues

4.3.4 Comparison Across Outcomes

The XGBoost model demonstrated only significant moderation in yielding accuracy when predicting the outcome of social media on mental health. The model was approximately 37% accurate in predicting when people felt depressed or down, 35% in predicting distraction, and 35% in predicting sleep-related problems, as shown in Figure 3. Overall, this model produces consistent results, but the performance is low. The findings suggest that social media use may have some quantifiable associations with mental health outcomes; however, the relationships are complex and likely influenced by other variables that are difficult to measure.

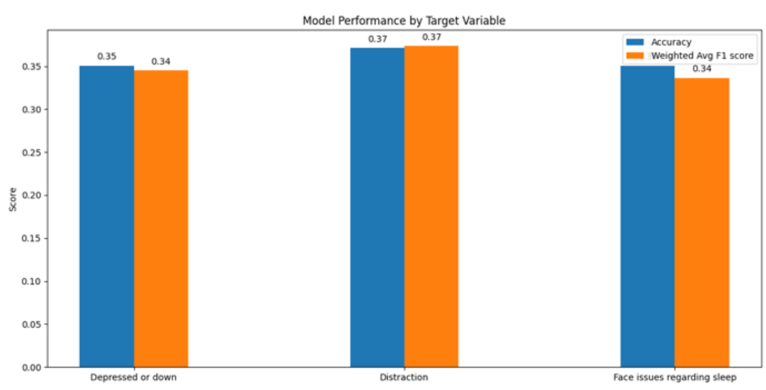


Figure 3. Model Accuracy and F1-Score by Target Variable (XG Boost Model)

ISSN: 2473-4901
v12 n6709

4.4 Multi-Output Classifier

Unlike the previous classifiers, the Multi-Output Classifier predicts binary outcomes for each target. For all three targets (depression, distraction, and sleep issues), Class 0 represents the absence of the condition (negative class), while Class 1 represents the presence of the condition (positive class).

4.4.1 Evaluating Model for Target: Depression

For the *Depression* target, the confusion matrix (see Table 19) shows strong diagonal dominance with minimal misclassification, indicating high predictive accuracy. As shown in Table 20, the Multi-Output Classifier model achieves a good result on this binary target, with an accuracy of 82%. Looking at Class 1, which has depressive symptoms, it is highly precise and recall, i.e., it is reliable in picking up individuals who are indeed depressed. Class 0 is also doing an acceptable job, only slightly worse with the recall. Overall, the model is very effective in separating cases where individuals are not depressed and those where they are probably depressed.

True \ Pred	*Class 0*	*Class 1*
Class 0	17	11
Class 1	6	63

Table 19. Multi-Output Classifier Model Confusion Matrix – Depression

Class	Precision	Recall	F1-Score	Support
0	0.74	0.61	0.67	28
1	0.85	0.91	0.88	69
Accuracy			0.82	97
Macro Avg	0.80	0.76	0.77	97
Weighted Avg	0.82	0.82	0.82	97

Table 20. Multi-Output Classifier Model Performance Metrics – Depression

4.4.2 Evaluating Model for Target: Distraction

For the *Distraction* target, the confusion matrix reveals that the errors made in the two classes are very minimal (see Table 21). The model achieves a high number of correct predictions with limited errors, demonstrating robust pattern recognition. As shown in Table 22, the model achieves an accurate at 87 % hence its ability to predict distractions is high. Class 1 is characterized by great accuracy and recall and is therefore highly effective in catching distracted individuals. Class 0 shows slightly lower recall compared to the others, but overall, its performance remains strong. Therefore, this model works remarkably well in the prediction of distraction in this binary configuration.

True \ Pred	*Class 0*	*Class 1*
Class 0	11	8
Class 1	5	73

Table 21. Multi-Output Classifier Model Confusion Matrix – Distraction

Class	Precision	Recall	F1-Score	Support
0	0.69	0.58	0.63	19
1	0.90	0.94	0.92	78
Accuracy			0.87	97
Macro Avg	0.79	0.76	0.77	97
Weighted Avg	0.86	0.87	0.86	97

Table 22. Multi-Output Classifier Model Performance Metrics – Distraction

4.4.3 Evaluating Model for Target: Sleep Issues

For the *Sleep Issues* target, the confusion matrix indicates good overall performance with some remaining misclassification, but still significantly better than the other models (see Table 23). The accuracy of the model is moderate, with only 68% accuracy for sleep-related problems, as shown in

 ISSN: 2473-4901
v12 n6709

Table 24. Predictions for Class 1 are strong, with high recall and good precision, meaning the model correctly identifies most individuals with sleep problems. In contrast, Class 0 has lower recall, indicating that some individuals without sleep issues are incorrectly classified as affected This imbalance is consistent with the confusion matrix. Overall, the performance of the model is moderate but not as high as in the case of the other two targets.

True \ Pred	*Class 0*	*Class 1*
Class 0	16	23
Class 1	8	50

Table 23. Multi-Output Classifier Model Confusion Matrix – Sleep Issues

Class	Precision	Recall	F1-score	Support
0	0.67	0.41	0.51	39
1	0.68	0.86	0.76	58
Accuracy			0.68	97
Macro avg	0.68	0.64	0.64	97
Weighted avg	0.68	0.68	0.66	97

Table 24. Multi-Output Classifier Model Performance Metrics – Sleep Issues

4.4.4 Comparison Across Outcomes

The Multi-Output Classifier model performed well in all the considered variables, including the negative impact of social media on mental health. The model achieved an accuracy of 82% for depression, 87% for distraction, and 68% for sleep-related issues as shown in Figure 4. This method, compared to the previous ones, was significantly improved in accuracy and reliability. These findings suggest that the Multi-Output Classifier is effective in capturing the multidimensional effects of social media on mental health, providing more robust predictions by integrating emotional, cognitive, and behavioral outcomes.

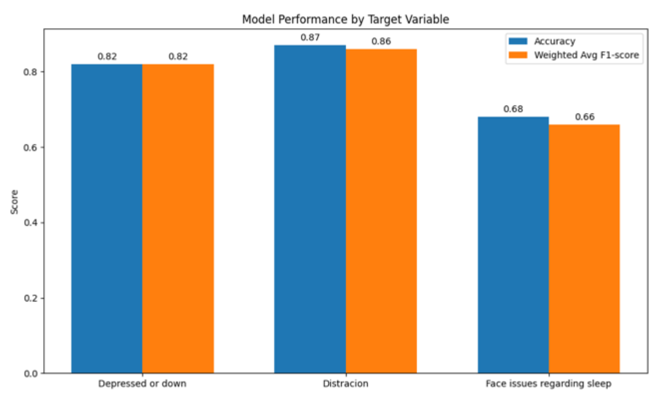


Figure 4: Model Accuracy and F1-Score by Target Variable (Multi-Output Classifier Model)

4.5 Comparison of All Models

The performance of four models, LR, SVM, XGBoost, and the Multi-Output Classifier, is compared across three mental health outcomes: feeling depressed or down, distraction, and sleep problems, as shown in Table 25 and Figure 5.

- Depression: The Multi-Output Classifier significantly outperforms other models (score ≈ 0.82), while LR, SVM, and XGBoost show low and comparable performance (0.37–0.40). This suggests that depression related effects of social media are complex and better captured by models that can consider multiple interrelated outputs simultaneously.
- Distraction: The Multi-Output Classifier again shows the highest performance (≈0.86), followed by SVM (≈0.56) and LR (≈0.50). This indicates that distraction is the most predictable outcome among the three, highlighting that social media usage has a strong and measurable association with attention-related impacts.
- Sleep Issues: All single-output models (LR, SVM, XGBoost) perform poorly (≈0.30–0.34), whereas the Multi-Output Classifier shows a higher score (≈0.67). This suggests that sleep-related effects

ISSN: 2473-4901
v12 n6709

are influenced by multiple interacting factors, which the Multi-Output model can capture more effectively.

Figure 6 shows the distribution of F1-scores across all models and targets, highlighting clear differences in performance. The Multi-Output Classifier is shifted toward higher F1-score values (around 0.65–0.85), indicating consistently better performance. In contrast, LR, SVM, and XGBoost are clustered in the lower range (approximately 0.30–0.55), reflecting weaker and less consistent results.

Overall, the Multi-Output Classifier model performs the best, and it confirms that the mental-health influence of social media is multi-dimensional and interlinked. The most directly measurable effect of social media is on distraction, whereas predicting depression and sleep issues requires a more holistic approach due to their complexity. In summary, the adverse effects of social media on mental health are subtle, impacting cognitive, emotional, and behavioral aspects in different ways.

Model	*Target*	*Precision*	*Recall*	*F1-score*	*Accuracy*
SVM	Depression	0.4	0.38	0.38	0.38
	Distraction	0.56	0.56	0.56	0.56
	Sleep Issues	0.3	0.31	0.3	0.31
LR	Depression	0.42	0.38	0.39	0.38
	Distraction	0.5	0.49	0.5	0.49
	Sleep Issues	0.28	0.31	0.29	0.31
XGBoost	Depression	0.38	0.37	0.37	0.37
	Distraction	0.35	0.36	0.35	0.35
	Sleep Issues	0.34	0.35	0.34	0.35
Multi-Output Classifier	Depression	0.82	0.82	0.82	0.82
	Distraction	0.86	0.87	0.86	0.87
	Sleep Issues	0.68	0.68	0.66	0.68

Table 25. Comparison on Performance across the Models

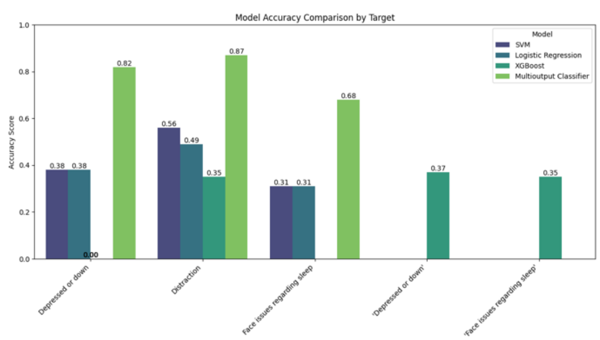


Figure 5. Model Performance Comparison by Target and Metric

 ISSN: 2473-4901
v12 n6709

Figure 6: Distribution of F1-scores across all Models and Targets

5. DISCUSSION

The primary objective of this study was to evaluate and compare different machine learning approaches for predicting mental health-related outcomes, including depression, distraction, and sleep issues, using social media usage data. The results demonstrate that model selection plays an important role in predictive performance. Among the evaluated models, including Logistic Regression, SVM, XGBoost, and the Multi-Output Classifier, the Multi-Output Classifier achieved the strongest overall performance across all three target variables. This indicates that jointly modeling multiple related outcomes can provide better predictive capability compared with treating each outcome independently.

A comparison of the evaluated models highlights the differences in their ability to capture patterns within the dataset. Logistic Regression provided a baseline approach by identifying linear relationships between input features and target variables. SVM and XGBoost improved prediction performance by capturing more complex patterns within the data. However, these models analyzed each target separately, which may limit their ability to consider relationships among related mental health outcomes. In contrast, the Multi-Output Classifier simultaneously considered depression, distraction, and sleep issues during prediction, allowing it to capture shared patterns among these outcomes and achieve improved performance.

The findings suggest that social media-related behavioral patterns may provide useful indicators for predicting potential mental health concerns. Machine learning models can assist in identifying patterns associated with increased risk and may support early identification of individuals who could benefit from additional attention or resources. However, applying such models in real-world settings requires careful consideration of privacy, data security, ethical concerns, and the potential consequences of incorrect predictions, particularly when analyzing sensitive mental health information.

Several limitations should be considered when interpreting the results. First, the dataset used in this study was relatively small and based on self-reported responses, which may introduce bias or inconsistencies due to differences in individual interpretation and reporting. Additionally, the dataset did not include more detailed behavioral information, such as the type of content users interact with, specific platform usage patterns, or time-based engagement trends, which could influence mental health outcomes. From a modeling perspective, limited hyperparameter tuning and the absence of external validation may affect the generalizability of the results to other datasets.

Future work should focus on evaluating these models using larger and more diverse datasets to improve reliability and generalization. Incorporating additional behavioral features, longitudinal data, and platform-specific usage patterns may provide deeper insights into social media-related mental health prediction. Furthermore, advanced machine learning approaches, improved validation strategies, and explainable artificial intelligence techniques could enhance model transparency and support responsible application in sensitive domains such as mental health analysis.

ISSN: 2473-4901
v12 n6709

6. CONCLUSION

This study evaluated and compared multiple machine learning models, including Logistic Regression, SVM, XGBoost, and a Multi-Output Classifier, for predicting mental health-related outcomes from social media usage data. The results demonstrate that the Multi-Output Classifier achieved superior performance across depression, distraction, and sleep issues, indicating the potential advantage of jointly modeling related outcomes rather than treating them independently. The findings suggest that machine learning approaches can effectively identify patterns associated with social media-related mental health outcomes. However, these results represent predictive associations within the dataset and should not be interpreted as evidence of causal relationships. The study highlights the importance of selecting appropriate modeling approaches for analyzing complex behavioral data. Future research should focus on larger and more diverse datasets, incorporation of additional behavioral features, longitudinal analysis, and explainable AI techniques to improve model generalizability, interpretability, and responsible application in mental health prediction.

7. REFERENCES

Ahmed, A., Aziz, S., Toro, C. T., Alzubaidi, M., Irshaidat, S., Serhan, H. A., Abd-Alrazaq, A. A., & Househ, M. (2022). Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update, 2*, 100066. https://doi.org/10.1016/j.cmpbup.2022.100066

Aichner, T., Grünfelder, M., Maurer, O., & Jegeni, D. (2021). Twenty-five years of social media: A review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, Behavior, and Social Networking, 24*(4), 215–222. https://doi.org/10.1089/cyber.2020.0134

Anand, U., Singh, R. K., & Gupta, K. (2024). Impact of social media on mental health. In *2024 International Conference on Communication, Control, and Intelligent Systems (CCIS)* (pp. 1–7). IEEE.

Aydın, A., Arda, B., Güneş, B., & Erbaş, O. (2021). Psychopathology of cyberbullying and internet trolling. *JEB Medical Sciences, 2*(3), 380–391.

Bashir, H., & Bhat, S. A. (2017). Effects of social media on mental health: A review. *International Journal of Indian Psychology, 4*(3), 125–131.

Bottino, S. M. B., Bottino, C. M. C., Regina, C. G., Correia, A. V. L., & Ribeiro, W. S. (2015). Cyberbullying and adolescent mental health: Systematic review. Cadernos de Saúde Pública, 31(3), 463–475. https://doi.org/10.1590/0102-311X00036114

Brown, T. A., Campbell, L. A., Lehman, C. L., Grisham, J. R., & Mancill, R. B. (2001). Current and lifetime comorbidity of the DSM-IV anxiety and mood disorders in a large clinical sample. *Journal of Abnormal Psychology, 110*(4), 585–599. https://doi.org/10.1037/0021-843X.110.4.585

Coyne, S. M., Rogers, A. A., Zurcher, J. D., Stockdale, L., & Booth, M. (2020). Does time spent using social media impact mental health? An eight-year longitudinal study. *Computers in Human Behavior, 104*, 106160.

Gao, J., Zheng, P., Jia, Y., Chen, H., Mao, Y., Chen, S., Wang, Y., Fu, H., & Dai, J. (2020). Mental health problems and social media exposure during COVID-19 outbreak. *PLOS ONE, 15*(4), e0231924.

Garg, M. (2023). Mental health analysis in social media posts: A survey. *Archives of Computational Methods in Engineering, 30*(3), 1819–1840.

Hernandez, J. M., Ben-Joseph, E. P., Reich, S., & Charmaraman, L. (2024). Parental monitoring of early adolescent social technology use in the United States: A mixed-method study. *Journal of Child and Family Studies, 33*(3), 759–776. https://doi.org/10.1007/s10826-023-02734-6

Hu, Y., Bai, Y., Pan, Y., & Li, S. (2021). Cyberbullying victimization and depression among adolescents: A meta-analysis. *Psychiatry Research, 305*, 114198. https://doi.org/10.1016/j.psychres.2021.114198

ISSN: 2473-4901
v12 n6709

Karim, F., Oyewande, A. A., Abdalla, L. F., Ehsanullah, R. C., & Khan, S. (2020). Social media use and its connection to mental health: A systematic review. *Cureus, 12*(6).

Kim, J., Lee, J., Park, E., Han, J., & Lee, H. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports, 10*, 11846. https://doi.org/10.1038/s41598-020-68764-y

Kuss, D. J., & Griffiths, M. D. (2011). Online social networking and addiction—A review of the psychological literature. *International Journal of Environmental Research and Public Health, 8*(9), 3528–3552.

Lamba, K., Rani, S., & Shabaz, M. (2026). Explainable machine learning for mental health prediction from social media behavior: A nested cross-validation study with SHAP and LIME interpretability. *Discover Mental Health, 6*, 25. https://doi.org/10.1007/s44192-026-00373-z

Liu, D., Feng, X. L., Ahmed, F., Shahid, M., & Guo, J. (2022). Detecting and measuring depression on social media using a machine learning approach: Systematic review. *JMIR Mental Health, 9*(3), e27244. https://doi.org/10.2196/27244

Naslund, J. A., Bondre, A., Torous, J., & Aschbrenner, K. A. (2020). Social media and mental health: Benefits, risks, and opportunities for research and practice. *Journal of Technology in Behavioral Science, 5*(3), 245–257.

O'Reilly, M., Dogra, N., Hughes, J., Reilly, P., George, R., & Whiteman, N. (2019). Potential of social media in promoting mental health in adolescents. *Health Promotion International, 34*(5), 981–991.

Tao, Z. (2024). Long-term use of social media is significantly associated with the prevalence of mental illness. Theoretical and Natural Science, 51, 194–198. https://doi.org/10.54254/2753-8818/51/2024CH0201

Shannon, H., Bush, K., Villeneuve, P. J., Hellemans, K. G. C., & Guimond, S. (2022). Problematic social media use in adolescents and young adults: Systematic review and meta-analysis. *JMIR Mental Health, 9*(4), e33450.

Sheng, N., Yang, C., Han, L., & Jou, M. (2023). Too much overload and concerns: Antecedents of social media fatigue and the mediating role of emotional exhaustion. Computers in Human Behavior, 139, 107500. https://doi.org/10.1016/j.chb.2022.107500

Starr, L. R., Davila, J., Stroud, C. B., Li, P. C. C., Yoneda, A., Hershenberg, R., & Miller, M. R. (2012). Love hurts (in more ways than one): Specificity of psychological symptoms as predictors and consequences of romantic activity among early adolescent girls. Journal of Clinical Psychology, 68(4), 403–420. https://doi.org/10.1002/jclp.20862

Vidal, C., Lhaksampa, T., Miller, L., & Platt, R. (2020). Social media use and depression in adolescents: A scoping review. *International Review of Psychiatry, 32*(3), 235–253. https://doi.org/10.1080/09540261.2020.1720623

Vilarinho-Pereira, D. R., Botelho, M., & Leal, C. (2021). Understanding the use of social media to foster student creativity: A systematic literature review. *Creativity: Theories – Research – Applications, 8*(1), 124–147. https://doi.org/10.2478/ctra-2021-0009