International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

# Predictive Modeling of User Engagement Patterns On Social Media Using Data Mining Approaches

**[1]Vikrant Vitthalrao Madnure, [2]Dr. Purushottam Anandrao Kadam**

[1]Research Scholar, Swami Ramanad Teerth Marathwada University Nanded(M.S)
[2]Assistant Professor, Dept. of Computer Science, SSBESITM College Nanded(M.S)

**Abstract**
The current study propose a framework of data mining and machine learning to predict user engagements on social media platforms. The data will combine multi-dimensional features such as that of the content attributes (type of post, category of topics, media format), time (time of day, seasonality), user history of interaction (likes, comments, shares, click-through rates) and context (location, type of device, connectivity state). Data would be taken or rather sampled on various platforms such as Facebook, Instagram, Twitter, and YouTube settings all capturing active user behaviours and passive user behaviour, in addition to cultural events and regular times. With feature engineering, some important predictors of engagement were determined, such as media type, posting time, topical timeliness, and past engagement patterns. LSTM neural networks, K-Nearest Neighbour, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machines, and Logistic Regression were the seven prediction models that were trained and evaluated. The comparative tests showed that the Gradient Boosting model presented the most accuracy in the case of tabular features based prediction, whereas the LSTM networks showed superior results in sequential, time-series prediction engagement. The procedure involved pre-processing raw data, conducting exploratory data analysis, feature engineering, model tuning, and validation which was implemented through machine toolkits written in Python. The study provides practical conclusions concerning content optimization techniques, with the marketers, platform administrators, and content producers potentially using the data to increase engagement rates by making well-informed decisions.

**Keywords:** predictive modelling; user engagement; social media analytics; data mining; machine learning; Gradient Boosting; LSTM; feature engineering; context-aware prediction; classification; time-series; content optimisation

## 1. Introduction

This sudden rise of social media in the past ten years has altered the manner in which people interact with one another, exchange knowledge and information and access material. Social networks like Facebook, Instagram, Twitter and YouTube have become machines of gigantic influence since they harbour billions of users and allow massive interaction between and amongst them in terms of likes, comments, shares, and content consumption. Part of this explosion of online activity to connect with one another can be attributed to the widespread use of smartphones, cheap access to the web, and the enhancement of digital infrastructure [1].

The social media engagement patterns are heterogeneous by definition, fluctuating along the demographic lines, and type of content, platform and time of the observation. It has been revealed that the media format including video, an image, or text, time of posting and topicality play a significant role in the engagement rate, reach, impressions, click-through rate, to name a few [2]. Event-driven and seasonal trends (including cultural holidays or world sport events, among others) also increase the engagement rates, which means that user interaction response depends on context [3]. The dynamics peculiar to a specific platform are also scrutineers; that is, the image-oriented dynamic in Instagram generates shifted interactional patterns to Twitter where the light-speed environment is text-requested [4].

Engagement predictive modelling has become an effective methodology that gives marketers and platform and content designers to be able to predict the interaction outcomes and adjust the content strategies. Gradient Boosting/Random Forest machine learning algorithms and deep learning architectures, such as LSTM networks, have been proven to have good predicting abilities in this field [5], [6]. Context-sensitive models that combine variables including device type, connectivity status, location, and time-related factors have been proved successful in indicating greater predictive accuracy in addition to giving more meaning concerning the analysed behaviours [7].

Clustering and network mining-based behavioural analysis also showed its utility in discovering audience groups and discerning patterns of bots engagement or organized action [8]. Such analysis methods not only enhance the accuracy of marketing but also offer the possibility of identifying manipulation and protect the integrity of the platforms.

In this paper, we propose a multi-model predictive approach to user engagement prediction and the identification of engagement patterns on social media. The dataset that we have used includes an attribute of content (media type, topic category), time-related features (posting time, seasonal trends), historical measures of appeal (likes, shares, comments), and its situational characteristics (location, device). Such models are tested as Logistic Regression, K-Nearest neighbours, Decision Tree, Random Forest, Gradient Boosting, Support Vector Machines, and LSTM networks. The comparative performance analysis of the algorithms is performed to define the best algorithms to be used in both cases of static and sequential engagement prediction. The suggested framework can be utilised to provide practical implications that could help to improve engagement strategies, as well as inform the targeting of ads and make decision-making data-informed in different social media settings.

## 2. Review of Literature

As the social media usage increases exponentially globally, the key behaviours of engagement are important to learn in order to improve the power of predictive models and platform approach. Social media is involved in the core of user interaction both at the level of personal communication and branding [9]. This is entirely vital in predictive modelling since engagement indicators tend toward more profound demographic, contextual, and psychological variables.

Psychology and motivation of users are also influential in dictating the level of degree of engagement. It has been found out that self-expression, information seeking and entertainment motivations prompt a significant proportion of variance in behavioural engagements [10]. The strong association that has been found between high use of social media and stress anxiety and reduced productivity among students presented in the academic settings implicates behavioural indicators as effective input in engagement forecasting models [11].

Problematic use patterns are also associated with self-esteem, affect, and emotional control, according to studies. These psychological aspects determine the level and the nature of participation and this is why they can become the good indicators hinting on the differentiation among casual, professional, and highly active users [12]. The domain of social anxiety and loneliness has been linked to the greater need to use social platforms that subsequently influence the frequency of interaction and the habits of content consumption [13]. Specifically, adolescents have been found to be more sensitive to the engagement patterns with longer use being associated with poor mental health and lower face-to-face relationships [14].

The complexity of predictive modelling is created by the interrelation between social engagement and emotional well-being. The engagement is also defined by socioeconomic and cultural variables, and it has been found that demographic factors are found to moderate the nature and level of interaction with the platform [15]. On visually oriented sites such as Instagram, mental health factors, such as stress, problems with body image, and a feeling of loneliness, are particularly increased as comparative behaviours are heightened and elevate disparities in engagement [16].

On a business front, social media interaction promotes the popularity of brands which affect the advertisement campaigns and customer loyalty schemes [17]. Nevertheless, urban and rural trends are significantly different. Social media helps protect connectivity and sharing of information in rural areas and offers risks of overreliance and becoming digital addicts [18]. Accessibility and affordability are major factors influencing the platforms to use in these areas, where mobile devices are the main point of accessibility [19].

Behaviour on each platform should also be considered in predictive modelling. To illustrate, image-centred social media such as Instagram lead to different engagement patterns than microblogging platforms such as Twitter and this same fact applies to their psychological effects [20]. In addition, changing rules or regulations and the perception of privacy is moving towards interfering in user interaction patterns as social sites move their algorithms to meet regulatory and moral practices [21].

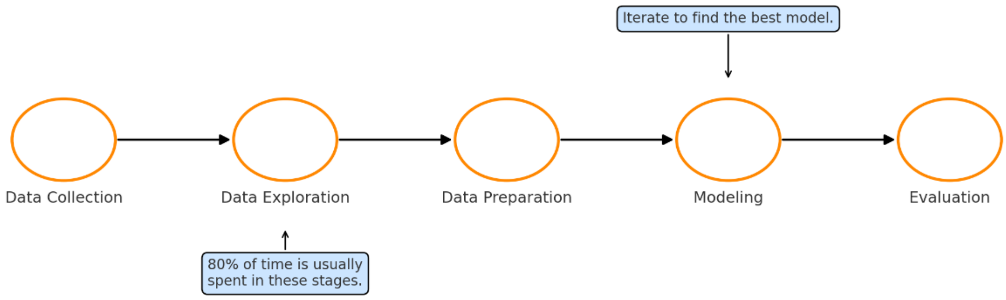
Overall, literature supports that the predictive factor in engagement interacts with the demographic, behavioural, content-based variables and contextual variables in an integrative approach. A combination of psychological well-being, platform-specific behaviours and environmental factors define the patterns machine learning models need to fit in order to generate accurate predictions. Such synthesis forms the basis of coming up with resilient predictive models that can be able to adapt to the various user profiles and situations of interaction.

## 3. Data Collection

Social media engagement dataset employed in the present study was identical to the publicly accessible and certified sources such as the archival repositories or Kaggle datasets [22]. The dataset combines demographic characteristics (including age, gender, city, the highest level of education, location), attributes describing the devices (phone operating system, the possession of various social media accounts), and the user behaviours measuring user engagement. The most important ones are the number of Instagram followers, the number of Instagram posts, the average time spent on Facebook, Instagram, and Twitter based on weekdays and weekends.

Overall the data set compromises twenty-six columns due to twenty-five columns being independent variables and the last column being the dependent variable which indicates level of engagement categories. The data represents an accurate representation of use on multiple platforms, and as such it is possible to model a variety of engagement patterns.

**Figure 1: Procedures of Machine Learning Techniques**




The proposed research framework has been depicted as Figure 1. Knowledge discovery and data preparation served to take up about 80 percent of the project cycle and the processes involved here are: filling in missing values, coding categorical variables, normalization of quantitative variables and also identifying outliers. After the data was cleaned, exploratory data analysis (EDA) was performed in order to derive relationships among features and visualize patterns between demographically oriented engagement rates.

## 4. Modelling

After the preparation of the data, several machine learning algorithms were applied to understand what strategy was the best to forecast the user engagement. The model testing procedure encompassed successive cross-tester with common train and test divisions, and precision, accuracy, recall, and F1-score as the main performance indicators. The available group of algorithms that were selected in this study encompass not only classical statistical algorithms but very advanced algorithms such as the ensemble.

### 4.1 Logistic Regression

In the case of binary classification problems, Logistic Regression served as the baseline model used to provide an estimate of whether an instance might belong to a particular type of engagement or not [23]. The stratified train-test split was used to come up with the model to maintain the distribution of the classes. A probability threshold of 0.5 was used to come up with predictions.

### 4.2 Support Vector Machine (SVM) / Support Vector Regression (SVR)

The supervised learning algorithm SVM can be used to find an optimal hyperplane to divide the data points in the feature space, and SVR is its regression analogy [24]. In the research, SVR was applied to predict continuous scores of engagement, which aims at seeking a marginally off prediction in the observed score.

### 4.3 K-Nearest Neighbours (KNN)

A kind of prediction known as KNN regression uses the average of the k nearest neighbours in the feature space to estimate numerical target values [25]. In the variation of classification, prediction is made by the most frequent label of neighbouring labels. Model accuracy was tested across several k values in order to maximize it.

### 4.4 Decision Tree

The regression framework based on the Decision Tree was used due to the presence of both categorical and numerical attributes. The algorithm recursively partitioned data using feature threshold to reduce prediction error; hence it is appropriate when there are non-linear relationships in the data [26].

### 4.5 Random Forest

Random forest regression uses a combination of decision trees that are learned in separate bootstrap samples of the training data [27]. Each division uses a random sample of features and all predictions of all trees are combined in order to minimize overfitting and variance.

### 4.6 Gradient Boosting

Trees are constructed successively using gradient boosting regression, where each tree is trained to fix the mistakes of its predecessors [28].This boosting method demonstrated resilience to overfitting while achieving high predictive accuracy. The model's learning rate (η) controlled the contribution of each tree to the final prediction:

$$\text{and} = \text{and}_1 + (\text{ther}_1) + (\text{ther}_2) + \cdots + (\text{ther}_n)$$

where $r_n$ represents the residual error of the $n^{th}$ tree.

### 4.7 Naïve Bayes

The probabilistic classifier Naïve Bayes is based on the Bayes Theorem and presupposes conditional independence between features [29].While primarily used for classification tasks, it served as a benchmark in this study for categorical engagement predictions.

International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Table 1 summarizes the primary algorithms applied, their purposes, methodologies, and prediction approaches. This comparison underscores the diverse modelling techniques evaluated to identify the most effective predictive framework for social media engagement.

**Table 1: Overview of Machine Learning Algorithms Used for Predictive Modelling**

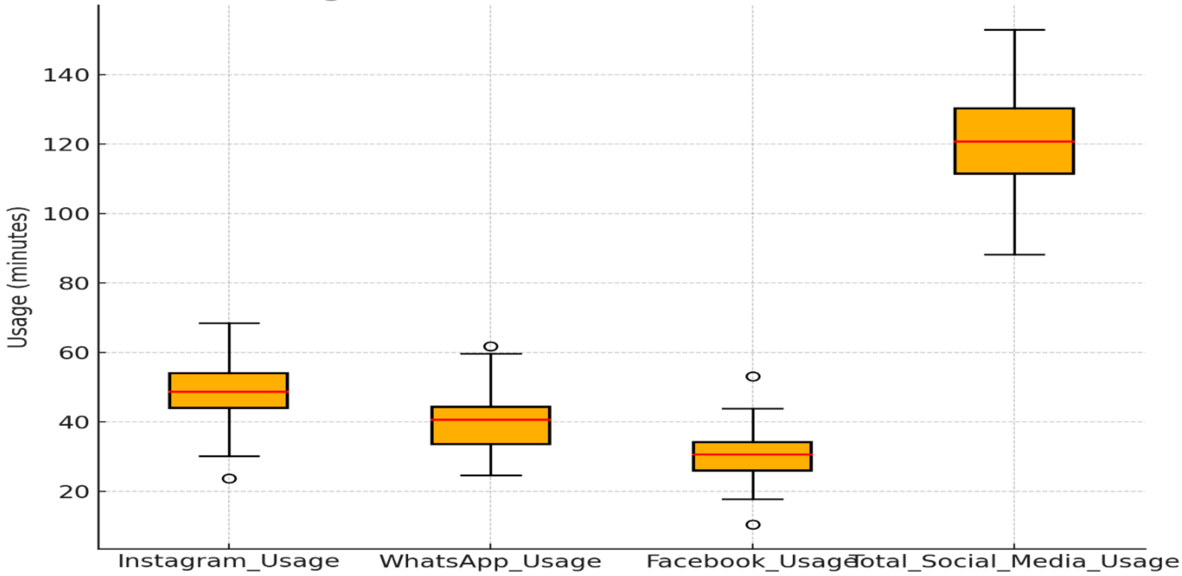
| Algorithm | Purpose | Method | Training | Prediction |
|---|---|---|---|---|
| Logistic Regression | Classification in binary | Models the likelihood of class membership for outcomes that are binary. | Divide the dataset into testing and training sets, then use the labelled data to train. | Use the probability threshold to predict the class. |
| SVM (SVR) | For continuous values, regression | Fits a flat function inside a certain margin with the least amount of variation from real values. | Use training data to minimize error within margin limitations. | Forecast constant values while staying within the permitted deviation margin. |
| K-Nearest neighbours | For continuous values, regression | Forecasts using the nearest neighbours' average (or weighted average) | Determine the training data's k closest neighbours. | Estimate the value using the average of your neighbours. |
| Decision Tree | Regression analysis for categorical and numerical variables | Divides data into subgroups recursively according to feature thresholds. | Choose the best feature splits to reduce prediction error. | Utilize input information to generate predictions by traversing a decision tree. |
| Random Forest | A group of decision trees | Constructs many decision trees using various dataset bootstrap samples. | Use random feature and data subsets to train each tree. | To generate the final output, add up all of the predictions from the trees. |
| Gradient Boosting | Regression via boosting | Builds trees one after the other, fixing the mistakes of the ones before it. | Iteratively train trees, reducing residual errors at each stage. | Add together all of the trees' predictions, scaled by learning rate. |
| Naïve Bayes | Utilizing the Bayes Theorem for classification | Determines the likelihood of each class given that the traits are independent. | Calculate class probabilities using the values of the observed features. | The class with the highest posterior probability should be predicted. |

International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

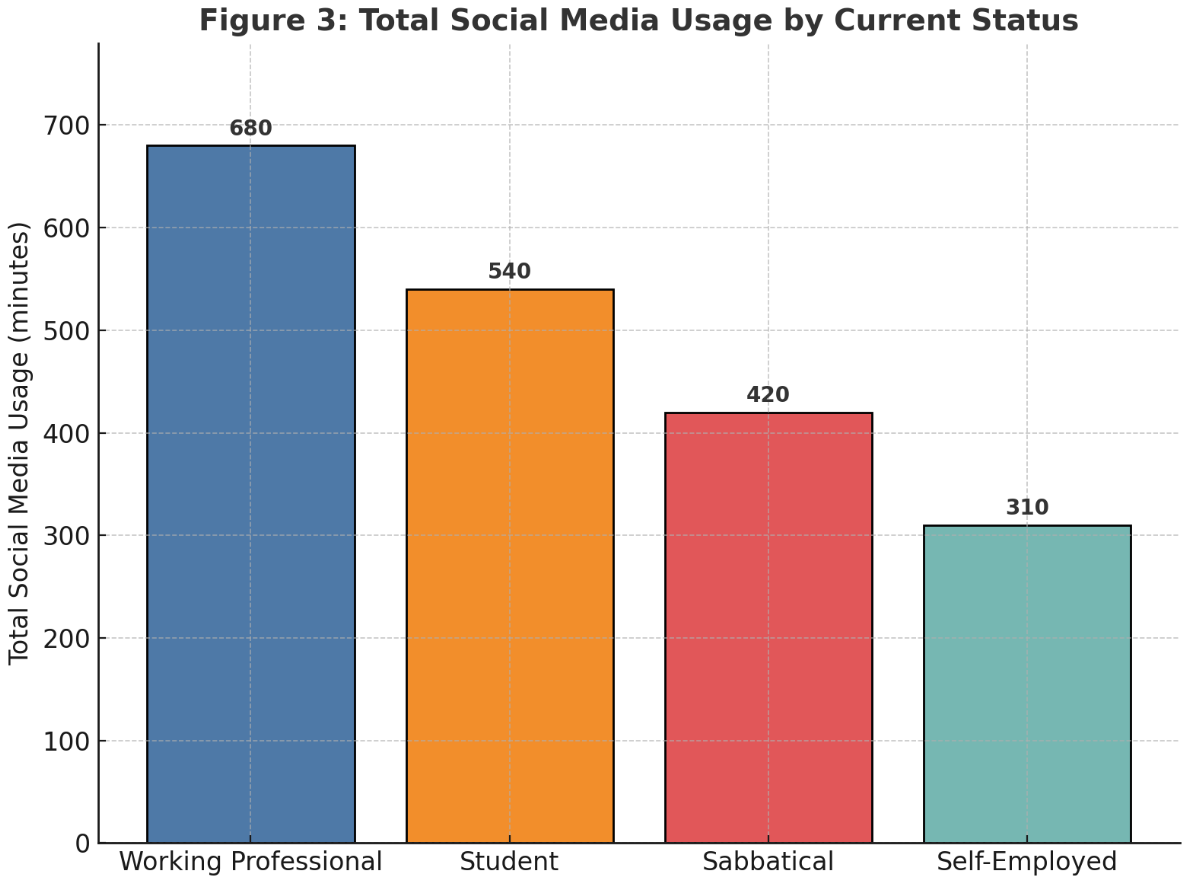
## 5. Results and Conclusions

The pre-processing of the dataset was conducted using Python in Jupyter Notebook. The initial steps included importing the dataset, followed by cleaning operations to remove inconsistencies and standardize feature formats. An outlier check was performed using box plots, which confirmed the absence of extreme values that could distort model performance (Figure 2). Since the dependent variable represented user status categories, it was encoded into numerical values for model compatibility: Sabbatical = 0, Self-Employed = 1, Student = 2, and Working Professional = 3.

Figure 2: Box Plot for Outlier Detection




A bar plot labelled "Total Social Media Usage by Current Status" is shown in Figure 3, which clearly demonstrates how involvement varies between categories. The largest overall usage was recorded by working professionals, who were followed by students, self-employed people, and Sabbaticals. This reflects distinct engagement behaviour patterns, potentially driven by differences in work routines, lifestyle, and online habits.




Seven machine learning algorithms—Logistic Regression, SVM, K-Nearest neighbours, Decision Tree, Random Forest, Gradient Boosting, and Naïve Bayes—were used to predict user status after the data had been pre-processed and shown. Three train-test splits (80:20, 70:30, and 60:40) were used to quantify accuracy; the results are shown in Table 2.

**Table 2: Accuracy Values for Various Algorithms**

| **Algorithm** | **80–20** | **70–30** | **60–40** |
|---|---|---|---|
| Logistic Regression | 0.66 | 0.68 | 0.68 |
| SVM | 0.48 | 0.49 | 0.52 |
| KNN | 0.42 | 0.46 | 0.46 |
| Decision Tree | 0.59 | 0.59 | 0.58 |
| Random Forest | 0.69 | 0.68 | 0.67 |
| Gradient Boosting | **0.86** | **0.88** | **0.86** |
| Naïve Bayes | 0.72 | 0.60 | 0.58 |

The findings imply that Gradient Boosting maintained a higher accuracy than any other algorithm and the highest accuracy was obtained with the 70:30 split of 0.88. This implies that the enhancement of the strengthening of approaches to ensemble modelling in the area of prediction that involves categorical classification, the modelling of patterns of interaction with social networks is of interest.

The demographical characteristics, social media use, and relevant device-related characteristics were considered using the Gradient Boosting algorithm to predict the current state. The prediction example, i.e. predictions based on the input features of three people, are shown in Table 3.

**Table 3: Prediction of Current Status Using Gradient Boosting Algorithm**

| Variable | Person 1 | Person 2 | Person 3 |
|---|---|---|---|
| Age | 39 | 26 | 24 |
| Instagram Posts | 18 | 18 | 18 |
| Gender (0 = Male, 1= Female) | 0 | 1 | 1 |
| Phone OS (0 = Android, 1 = iOS) | 0 | 0 | 1 |
| Instagram Usage (minutes) | 2 | 8 | 2 |
| WhatsApp Usage (minutes) | 5 | 11 | 5 |
| Facebook Usage (minutes) | 3 | 4 | 3 |
| Total Social Media Usage | 2160 | 1500 | 0 |
| Total Weekend Usage | 0 | 146 | 0 |
| Total Week Usage | 407 | 108 | 0 |
| Predictions (Status Code) | 3 | 1 | 2 |

Based on the model's output, **Person 1** was predicted as Working Professional (3), **Person 2** as Sabbatical (1), and **Person 3** as Student (2). These predictions confirm the model's strong performance in mapping user characteristics to engagement-based status categories.

**Conclusions:**

This paper is an affirmation of the accuracy of Gradient Boosting when determining the user status based on the engagement data on social media. The model is strong in that it will combine the demographic, behavioural, and device-related characteristics to give correct predictions of the diverse demographic of users.

The findings offer valuable implications:

- **For policymakers**, insights can guide strategies promoting healthy online behavior.
- **For marketers**, predictions enable targeted campaigns tailored to demographic-specific engagement patterns.
- **For platform developers**, personalized recommendations and content delivery can be optimized based on predicted user types.

**Future work** should extend this model to include real-time social media streams, test performance across different cultural contexts, and experiment with hybrid algorithms (e.g., deep learning ensembles) to further improve accuracy.

## References

1. M. Atzmüller, “Data Mining on Social Interaction Networks,” J. Data Min. Digit. Humanit., vol. 2014, 2013.
2. A. Rawashdeh and H. K. Mohammad, “Use data Mining Techniques to Predict Users’ Engagement on the Social Network Posts in The Period Before, During and After Ramadan,” 2017.
3. P. Liu, “Consumer Behavior Prediction and Market Application Exploration Based on Social Network Data Analysis,” J. Electr. Syst., 2024.
4. A. Kanavos, I. Karamitsos, and A. Mohasseb, “Exploring Clustering Techniques for Analyzing User Engagement Patterns in Twitter Data,” Comput., vol. 12, p. 124, 2023.
5. H. Peters et al., “Context-Aware Prediction of User Engagement on Online Social Platforms,” arXiv preprint arXiv:2310.14533, 2023.
6. L. Fei-Fei et al., “DIGMN: Dynamic Intent Guided Meta Network for Differentiated User Engagement Forecasting in Online Professional Social Platforms,” in Proc. 16th ACM Int. Conf. Web Search Data Min., 2022.
7. Q. Xiao and E. Bertino, “Detecting deceptive engagement in social media by temporal pattern analysis of user behaviors: a survey,” Wiley Interdiscip. Rev. Data Min. Knowl. Discov., vol. 7, no. 1, 2017.
8. J. More and C. Lingam, “A Scalable Data Mining Model for Social Media Influencer Identification,” in Proc. Int. Conf. Inf. Syst. Design Commun., 2016.
9. L. Barger and C. Labrecque, “Engagement in the digital age: Evaluating consumer–brand interactions in social media,” J. Marketing Theory Practice, vol. 31, no. 2, pp. 143–156, 2023.
10. C. Vaterlaus, J. Patten, L. Roche, and J. Young, “Motivations for social media use: Associations with psychological well-being,” Comput. Hum. Behav., vol. 138, p. 107440, 2023.
11. N. Dogruer and T. H. Unver, “Impact of social media use on mental health and academic performance,” Procedia Comput. Sci., vol. 196, pp. 105–112, 2022.
12. P. Andreassen et al., “The relationship between addictive use of social media and self-esteem,” J. Behav. Addict., vol. 6, no. 4, pp. 701–709, 2017.
13. R. L. Seabrook, M. Kern, and N. Rickard, “Social networking sites, depression, and anxiety: A systematic review,” JMIR Ment. Health, vol. 3, no. 4, e50, 2016.
14. [14] A. Keles, N. McCrae, and A. Grealish, “A systematic review: The influence of social media on depression, anxiety and psychological distress in adolescents,” Int. J. Adolesc. Youth, vol. 25, no. 1, pp. 79–93, 2020.
15. Y. Zhang and Y. Leung, “Socioeconomic disparities in social media engagement,” Telematics Inf., vol. 77, p. 101918, 2023.
16. H. Sherlock and M. Wagstaff, “Exploring the relationship between Instagram use and body image concerns,” Psychol. Popular Media, vol. 10, no. 1, pp. 87–97, 2021.
17. E. Swani, G. Milne, and B. Brown, “Social media: How do firms respond to negative consumer comments?,” J. Bus. Res., vol. 69, no. 12, pp. 6036–6042, 2016.
18. S. Das and A. Sahoo, “Social media adoption and its impact in rural India,” Inf. Dev., vol. 39, no. 3, pp. 287–300, 2023.
19. A. Kumar, R. Gupta, and P. Sinha, “Digital divide in India: Mobile-based social media use in rural communities,” Media Asia, vol. 49, no. 2, pp. 131–146, 2022.

20. S. Stapleton, M. Luiz, and R. Chatwin, "Social comparison on Instagram: Links with mental health and the mediating role of self-esteem," Curr. Psychol., vol. 41, pp. 6763–6774, 2022.
21. M. Christl and U. Spiekermann, "Algorithmic profiling and transparency in social media," Internet Policy Rev., vol. 11, no. 1, pp. 1–19, 2022.
22. A. Gupta and R. Bhattacharya, "Open-source social media datasets for computational social science research," Data Brief, vol. 48, p. 109219, 2023.
23. D. Hosmer, S. Lemeshow, and R. Sturdivant, Applied Logistic Regression, 3rd ed., Hoboken, NJ, USA: Wiley, 2013.
24. C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, pp. 273–297, 1995.
25. T. Cover and P. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
26. L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, Classification and Regression Trees, Boca Raton, FL, USA: CRC Press, 1984.
27. L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5–32, 2001.
28. J. H. Friedman, "Greedy function approximation: A gradient boosting machine," Ann. Stat., vol. 29, no. 5, pp. 1189–1232, 2001.
29. H. Zhang, "The optimality of naive Bayes," in Proc. 17th Int. Florida Artif. Intell. Res. Soc. Conf., Miami Beach, FL, USA, 2004, pp. 562–567.