Journal of Media Horizons
ISSN: 2710-4060 2710-4052 Volume 6, Issue 3, 2025

DATA-DRIVEN PREDICTION OF USER ENGAGEMENT ON FACEBOOK USING PYTHON AND ML ALGORITHMS

Abeer Shahzad[1], Ali Mujtaba Durrani[*2], Muhammad Uzair[3], Laiba Gul[4]

[1]*Department of Journalism and Mass Communication, University of Peshawar, Pakistan*
[*2]*Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, Pakistan*
[3,4]*Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Pakistan*

[1]abeershahzad123@gmail.com, [*2]ali@cecos.edu.pk, [3]muhammaduzair98@gmail.com, [4]laibalg6117@gmail.com

DOI: https://doi.org/10.5281/zenodo.15796304

Keywords
Social Media Engagement, Facebook, Journalism, Prediction, Machine Learning, Python

Article History
Received on 26 May 2025
Accepted on 26 June 2025
Published on 02 July 2025

Corresponding Author: *
Ali Mujtaba Durrani

Abstract
The audience engagement in the modern digital media environment is an important question that allows attractively calculating its strategy in content and the efficiency of communications. This paper describes a machine learning model written in Python that attempts to predict the interaction of user with posts on Facebook which is represented in terms of number of likes, shares, and comments. Based on a well-organized data of Facebook performance indicators, we used data preprocessing and feature engineering methods and we trained and compared various predictive models. Specifically, four regression algorithms were tested including XGBoost with hyperparameter tuning, LightGBM, Random Forest, and Gradient Boosting on the potential to model the engagement behavior. The results indicate powerful predictive abilities of ensemble learning algorithms, especially predictive capabilities of analyzing patterns leading to social media reaction. This strategy proves that data science can be used in journalism and mass communication and give media professionals tools they can acting upon to better analyze their audiences, plan their editorial calendars, and execute strategic distribution of their content. The given framework can be entirely implemented in Python and transferred to other digital platforms.

INTRODUCTION

With the advent of the digital age, news organizations can no longer think of their work on social media platforms, since the latter served as the standard method of distributing journalism to the audience (Ali, 2023). The COVID-19 pandemic even boosted this tendency by having demonstrated the importance of social media in relaying news and building the discourse concerning it (Alhuntushi, 2020). But to successfully maneuver thru these forums, more than mere intuitively based posts is necessary, they will need data based knowledge into the factors that spur engagement. It has been demonstrated in recent works that commonly used machine learning models, including XGBoost, LightGBM and Random Forest, are also useful in modeling the non- linear dependencies in social media data and can achieve superior performance to the traditional regression approaches (Carta, 2020). AI is being used by journals to provide data diagnostics, fact-checking and modeling the behavior of the audience (Number Analytics, 2025). However, this is the area lacking implementation of these models to specifically predictive analytics of encountering audience in Facebook post or specifically by academic research

studies in mass communication. Such one work has been done on social media news trends (Shahzad, 2025). Where traditional journalism was based on editorial feel and historical trends of the audience, the volatile and quick changing environment of social media requires interpretation of data in real-time. Tandoc and Maitra (2021) note that in light of the emergence of analytics-influenced journalism, newsrooms themselves have been redefined to such an extent that they are now often driven by the information contained in data on user clicks, shares, and comments, which is actively used to make decisions on what, and in what manner to write. It is not just that social media like Facebook is a way to distribute the news but also a feedback loop where journalists can get an idea of what people think and how they will behave at large scale (Kalsnes, 2022). It is, therefore, not only a technological issue to know how user engagement works, but rather becomes a strategic need by news organizations that want to get visibility and influence in what is becoming an increasingly dense digital arena.

1. Literature Review

Although the current demand on data is on the rise, it is clear that most newsrooms do not have the technical facilities or experience to successfully deploy predictive analytics (Ferrer-Conill & Tandoc, 2021). Engagement strategies mostly lack predictability, which is required to implement proactive content plans, especially since most plans are only descriptive and concerned with retrospective measures. A way forward is available, through the newer advancements in open-source tools, especially those related to machine learning libraries written in the Python language, to make access readily available to predictive modeling to both scholars and practitioners of communication (Ali & Tajuddin, 2023). By combining these tools with powerful modeling tools, this work has contributions as much to the technical literature in regards to the performance of regressions as much as it has to the newer discipline of computational journalism, where technology is used to improve the editorial process, not to substitute it. A summary of the related work is depicted in table 1:

Table 1: Summary of Related Work on Social Media Engagement Prediction

Author(s) & Year	Focus of Study	Method / Model Used	Research Gap
Carta et al. (2020)	Engagement prediction on Facebook using multimodal features	Deep neural networks and gradient models	Lacks application to journalism; focused on corporate content marketing
Ferrer-Conill & Tandoc (2021)	Audience metrics in digital newsrooms	Qualitative ethnographic study	No predictive modeling; focused on perception, not outcome prediction
Elkalliny (2021)	Facebook as a health communication tool during COVID-19	Descriptive engagement metrics	Domain-specific (health); lacks predictive machine learning framework
Jawley & Fahmy (2022)	Impact of social analytics on global journalism content decisions	Survey-based audience perception analysis	Focused on newsroom culture; no empirical modeling or algorithmic engagement
Kalsnes (2022)	Influence of platforms on journalism distribution	Policy and platform analysis	Explores platform power but lacks data modeling or performance evaluation

Author(s) & Year	Focus of Study	Method / Model Used	Research Gap
Ali & Tajuddin (2023)	Python-based ML tools for journalists	General ML guidance using Python	Lacks benchmarking of models or quantitative performance evaluation
Keco et al. (2024)	Predicting engagement on university Facebook image posts	RF, SMO, J48 classifiers in WEKA + visual feature engineering	Focus on image posts; lacks temporal engagement modeling and broader content types
Arazzi et al. (2023)	Predicting tweet engagement using graph neural networks on Twitter	Graph Neural Networks (TweetGage)	Platform-specific (Twitter); not Facebook or posts with multimedia/content metadata
Obucic et al. (2023)	Engagement prediction for university Facebook image posts	J48, SMO, RF classifiers with Vision API features	Similar to Keco et al.; image-only focus with no feature tuning or log-scale predictive modeling
Our Work (2025)	Predicting Facebook post engagement for journalism using ML and Python	XGBoost, LightGBM, RF, GBM with GridSearchCV	Fills the gap by applying, comparing, and optimizing ML models for journalism

The research paper helps fill the conceptual gap by providing a workable and replicable framework of modeling Facebook interaction in terms of likes, share, and comments using Python and ensemble machine learning strategies. In order to improve model accuracy and data robustness, we use log-transformations and improvements to feature engineering. We can add to that by benchmarking four of the state-of-the-art regression models XGBoost with model hyperparameter tuning using GridSearchCV, LightGBM, Random Forest, and Gradient Boosting using a dataset of 500 posts of the Facebook and 19 predictive variables. With the focus on both deployability and interpretability, this paper can give journalists and media strategist a practical and convenient tool that can be used to best optimize content format, time-of-publication, and audience targeting, in active newsrooms.

2. Methodology

This paper gives a quantitative and machine learning approach to foretell the user interest on Facebook posts where structured data and ensemble regression models are used. The panel is a set of data acquired via UCI Machine Learning Repository [12] and contains 500 posts of a mass cosmetics brand with 19 variables, such as post metadata, the time of publication, page indicators, and likes, shares, and comments as the engagement indicators. The characteristics form the balanced foundation of the interpretation of the impact of content and timing upon the interaction with the audience. The figure 1 presents the full methodology that has been applied in the study. It comprises the data retrieved in UCI repository, data cleansing as log transforms and feature engineering, model selection and training with Ensemble of Models (XGBoost, LightGBM, Random Forest, and Gradient Boosting), and assessment on indicators such as MAE, RMSE and R2 Score. It runs the process in the Python language, and it is both reproduction and scaled to be applied in journalism and digital media analytics.

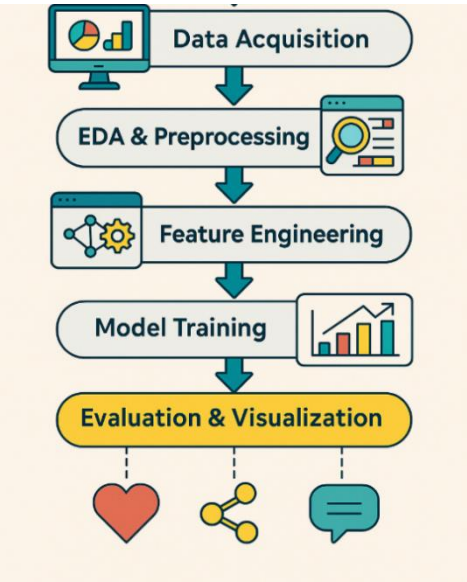

Figure 2: Methodology Process of Facebook Engagement Prediction by Machine Learning

In order to model the data, we dropped rows that contained null values in important columns (like that of Paid promotion status and engagement variables). Then feature engineering was performed to generate new predictors. As an example, we added two new variables: so-called Interaction_per_like = Total interactions divided by Number of likes (plus one to avoid divide by zero), and Paid_Hour_Interaction = (Paid promotions + Posting hour) - 1. In order to deal with the non-proportional nature of the engagement count, we log transformed results (using the natural log of one plus the value (log1p)) in like, share, and comment variables. Besides, the time-related characteristics like Post_Hour were merged in the form of the categorical buckets on the basis of the peculiarities of Facebook user behavior, e.g. Morning, Afternoon, Evening, and Night. We used four types of machine learning models in the process of engaging prediction: XGBoost (with the hyperparameter optimization based on the GridSearchCV), LightGBM, Random Forest, and Gradient Boosting. Such ensemble-based regressors were selected because they are robust, interpretable and excel at dealing with non-linear relationships and analysis of mixed data types. Models were all conducted through scikit-learn, XGBoost and LightGBM libraries of the Python open-source ecosystem. A typical training 80/20 of the data was used, and 3-fold cross-validation was used in GridSearchCV of the hyperparameters in the XGBoost model. Model evaluation pertained to three parameters namely Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R2 Score). Every model was trained and tested individually on three targets (log-transformed likes, shares, and comments). This way enabled us to compare the predictive power of various algorithms under various numbers of forms of audience engagement to easily come up with an overall picture of what motivates people to interact with Facebook content.

In the current project, data preprocessing, modeling as well as evaluating tasks were all performed via the use of Python as the language is very flexible and has an extremely powerful data science support system. The investigation was based on a number of open libraries. Data loading, cleaning, and feature engineering were performed with the help of pandas library, whereas the logarithmic transformations were performed with the help of numpy. As a means of data visualization and interpretation, matplotlib and seaborn have been used to create distribution plots, boxplots, heatmaps and charts demonstrating the differences of models. Machine learning models were used with the help of scikit-learn; the latter delivered the basis of model training, cross-validation, and model performance testing. XGBoost and LightGBM

libraries were built to create advanced ensemble models because they scale well and produce good results on regressions. To access the dataset programmatically we have used ucimlrepo library which retrieves the dataset directly at the UCI Machine Learning Repository. All the processing was made either in Jupyter Notebook or Google Colab environment, which makes it easy to reproduce. To optimize the model, especially the XGBoost one, we applied GridSearchCV imported in scikit-learn to conduct hyperparameter tuning with 3-fold cross-validation. It systematically tries the different combinations of hyperparameters to determine the combination that gives the best model performance on the basis of R2 score. The hyperparameters were n_estimators (number of boosting rounds) with the setting of 100 and 200; and max_depth (set at 3, 5, and 7 to regulate tree complexity and avoid overfitting) and learning_rate (where it was set to 0.05, 0.1, and 0.2, to balance the speed of training and generalization). R2 score was used as an evaluation metric of tuning and the random state of 42 was fixed so that it could be repeated. Such methodological configuration allowed to have an exhaustive and scalable architecture that could be applied not only in technical replication but also in an adaptation of newsroom conditions in which user engagement metrics are increasingly instrumental in editorial and content strategy-making processes.

3. Insights and Descriptive Statistics

In this research, the utilized dataset includes Facebook posts amounting to 500, with 19 variables, few of them are time, promotion, kind of content and the engagement. There is also a significant skewness in the distribution of the engagement metrics where likes varied between 0 a 5,172, shares between 0 a 790, and comments between 0 a 372. The median is 101 and the mean is 178 which means that a few posts that are very viral push up the mean. This was the reason that logarithmic transformation (log1p) was used to stabilize the variance and enhanced the model performance. Photo-type contents are prevalent in the posts (426 out of 500 posts) with a clear biased preference toward visual media. Paid variable displays that posts were sponsored only by ~28%, but the presence of this variable in prediction of engagement had a crucial impact, in the event alone and, in combination with the timing features, as Post_Hour was used. The Page_total_likes is between 81,370 and 139,441, signifying a high number of audience base and this is what rids of the large disparity in the reach and engagement.

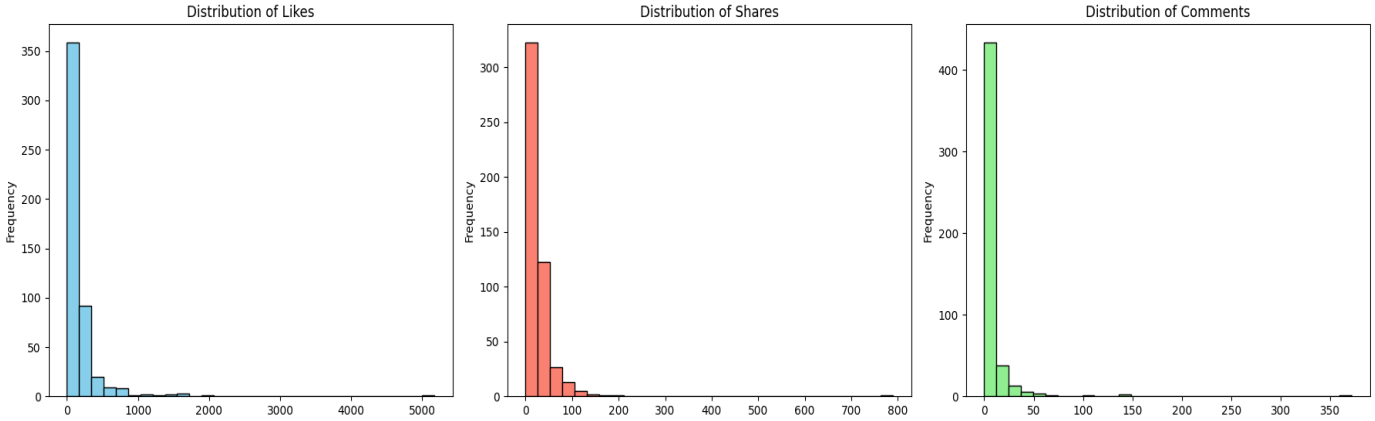

Figure 3. Shares, Likes, and Comments Given Out (Raw Scale).

It is worth noting that, the Lifetime_Post_Total_Reach and Lifetime_Post_Total_Impressions are highly spread with maximum values going up to 1.1 million impressions, which underlines the necessity of use of strong models able to cope with outliers. Other post, such as Post_Hour (mean 8 AM 19), and Post_Weekday (mean mid-week 19) showed some interesting trends and were later mapped into buckets in the behavioral insight. The pattern of engagement metrics namely liking, sharing and commenting was analyzed to identify the variation and distribution of

engagement of Facebook posts among the users. In Figure 3, the three metrics are significantly skewed to the right, which implies that most posts achieve a moderate level of interaction, whereas some viral posts are there as the outliers. As an example, of all the posts there is a huge majority with less than 200 likes and a minor part with at least 5,000 likes. This disparity is even greater when it comes to shares and comments and most of the posts have less than 50 shares or comments. The disparity justifies the application logarithmic transformation (log1p) to normalize target variables before the techniques of the regression modeling, hence enhancing the performance of the models and lessening the effects of the extreme scores.

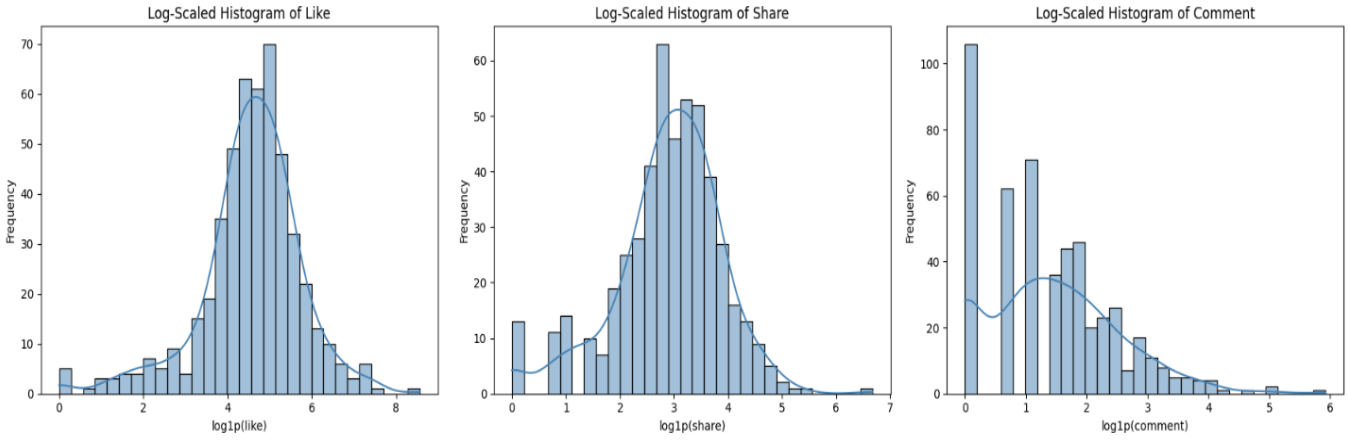

Figure 4. Log-Scaled Histograms of Likes, Shares and Comments.

The log1p transformation assists in the reduction of skew to the engagement measures and enhances symmetry to support better regression modeling. The like, share, and comment variables had raw engagement data that were log-transformed using log1p() to get rid of the non-normal distribution and extreme outliers. Although the distributions based on the log-scale presented in Figure 4 are rather normal and symmetrical, particularly when it comes to likes and shares. The transform can enhance the caliber of regression models through stabilizing variance and Gaussian sing error distributions. Despite the fact that, even after transformation, the distribution of comments is a bit skewed towards the right side of the scale, there is still a significant improvement in the basis of the raw scale. This preprocessing action was crucial in the aspect of training suitable models and ensuring similarity in performance on diverse algorithms. Regarding the behavior context and temporal dynamics of Facebook post engagement, we engaged in an analysis of post types, categories, days of week, and post hours distributions. The large proportion of photo posts in the dataset shows that there is a great preference in visual content as shown in Figures 5 , 6 ,7 and 8. The most common posts were rated as type 1, followed by category 3, 2. The analysis of weekdays was quite balanced, but activity was slightly higher at weekends (days 6 and 7). Looking at the hours the posts were posted, most of it was done in the early hours especially at 3 a.m. and the busiest time on this was the 10 a.m. and noon. These temporal and categorical insights are essential to journalists and media planners who wanted to streamline the audience reach/engagement by matching the time and type of content with the behavior of users.

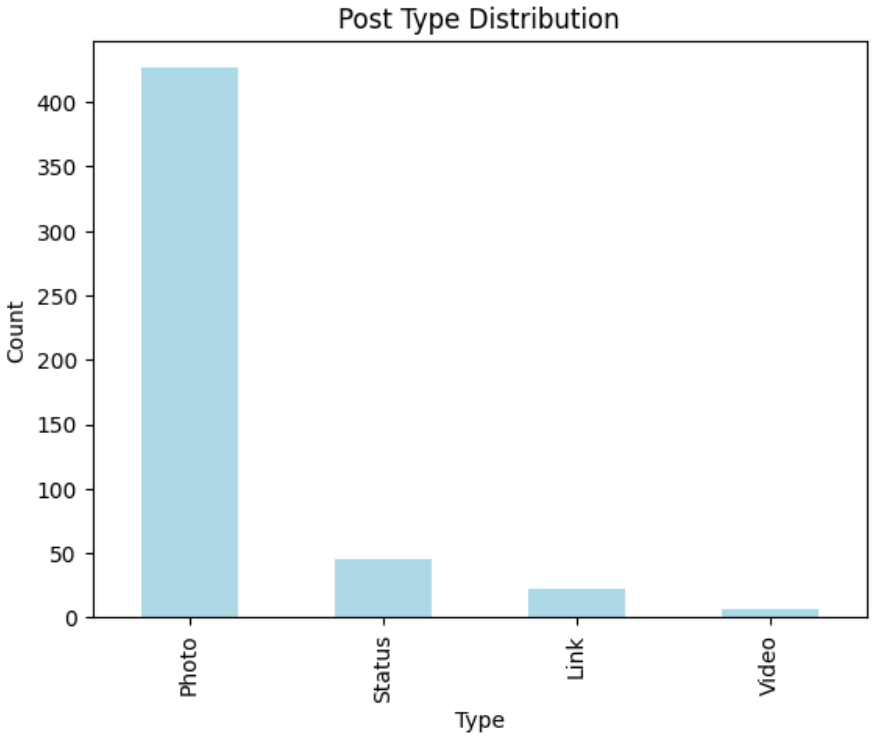

Figure 5. Bar plot of Distribution of Post Type.

The number of photo-based posts has been observed to prevail the most in most of the samples in the Facebook dataset followed by status posts, link and video posts which suggest dominance of the visual content.

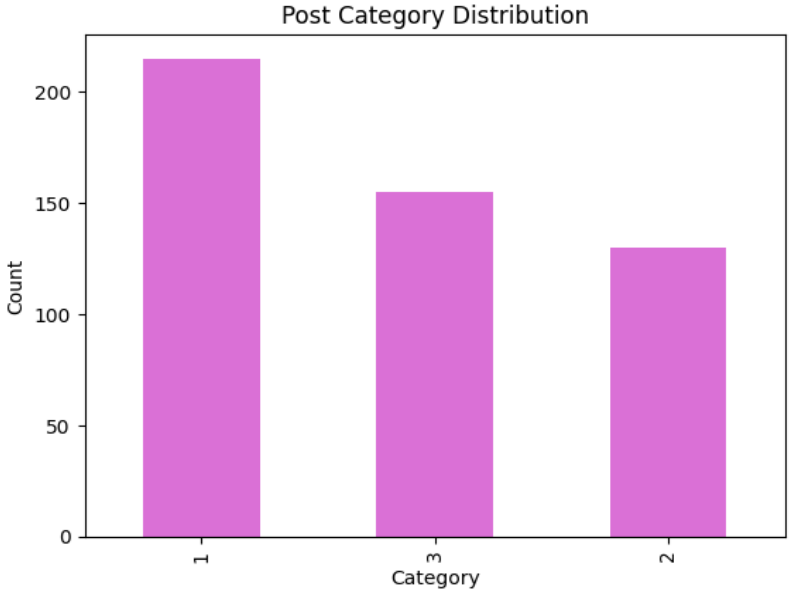

Figure 6. Bar plot of Category Distribution of Posts.

The most popular post category is Category 1 and the next two popular categories are Categories 3 and 2 that can be discussed as the topical type of the content.

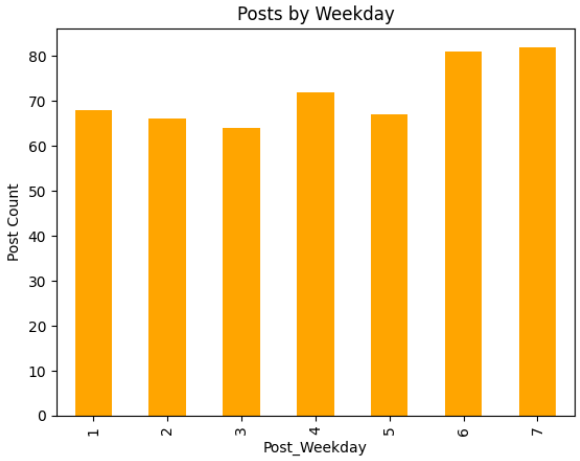

Figure 7. Bar Plot of Weekday Posts.

The distribution of posts throughout the week is rather equal, and increased frequencies are observed during weekends (days 6 and 7). There is a marked posting concentration at around 3 a.m., 10 a.m and noon, which could be the availability time of the audience and perhaps automation of the posting procedures. Figure 8 shows the correlation matrix offers a great deal of information as to the relation of various Facebook measures to each other. There is a high positive coefficient when it comes to the relationship between core engagement indicators, such as likes, shares, comments and Total_Interactions, often more than 0.85, establishing a relationship of dependency among them as important user interaction indicators. The lifetime measures also are highly strongly correlated (more than 0.7) with engagement measures, Lifetime_Engaged_Users,
Lifetime_Post_Consumptions are also equally much useful in modeling purposes. On the other hand, softer individual effects are observed in post-level temporal features such as Post_Hour and Post_Weekday since they have little correlation with engagement. This was used to to determine high-impact features to be used in further machine learning models.

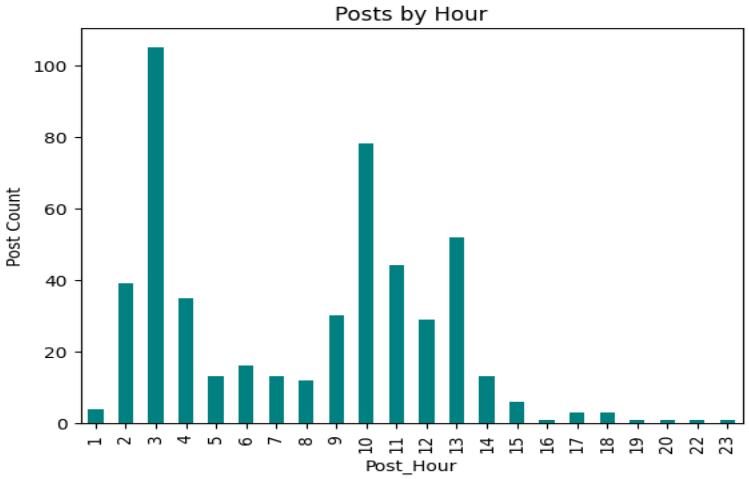

Figure 8. Bar Plot of Searched Based on Posts by Hour.

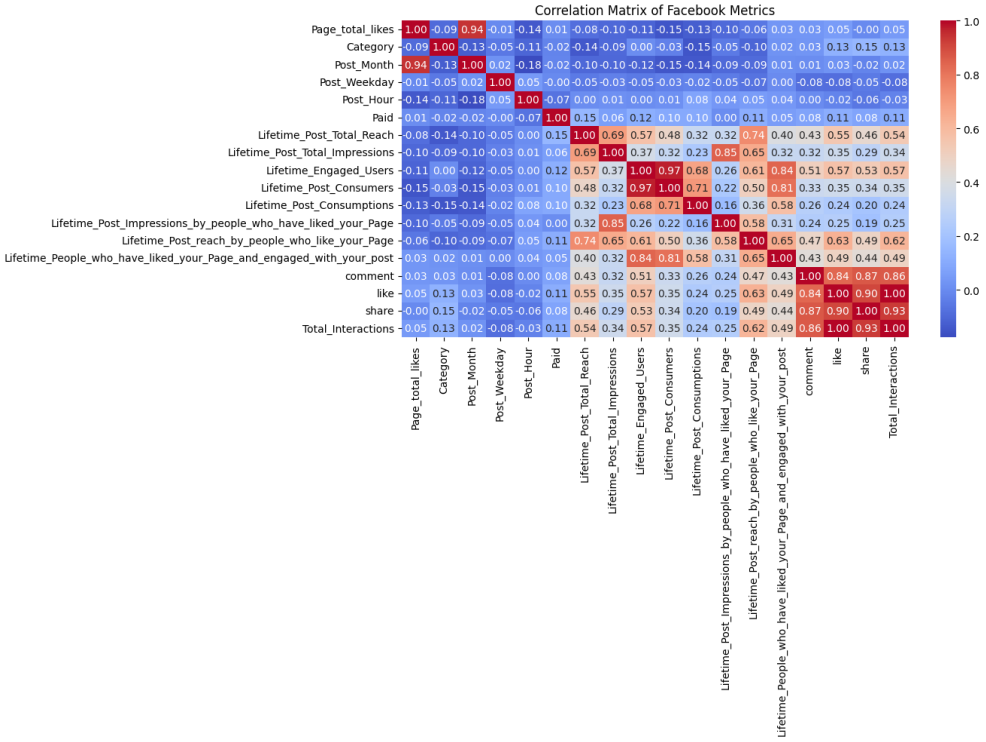

Figure 9. Correlation Matrix of Facebook Metrics.

The heatmap indicates the high levels of the positive correlations present within the engagement-related metrics and includes Lifetime_Engaged_Users and Lifetime_Post_Consumptions as the main predictors.vIn order to go even deeper into understanding of engagement behavior by time and content category, series of heatmaps and compare plots were created. It is revealed that engagement (likes, shares, comments) would differ greatly based on hour of the day, as well as day of the week and can therefore be used between time and days to ensure that the optimal number of users are engaged (Figures 10 to 12). Particularly, the average values of all of the three performance measurements are the greatest when the post is posted in the early morning time of day, particularly in and around Wednesday 5 AM.

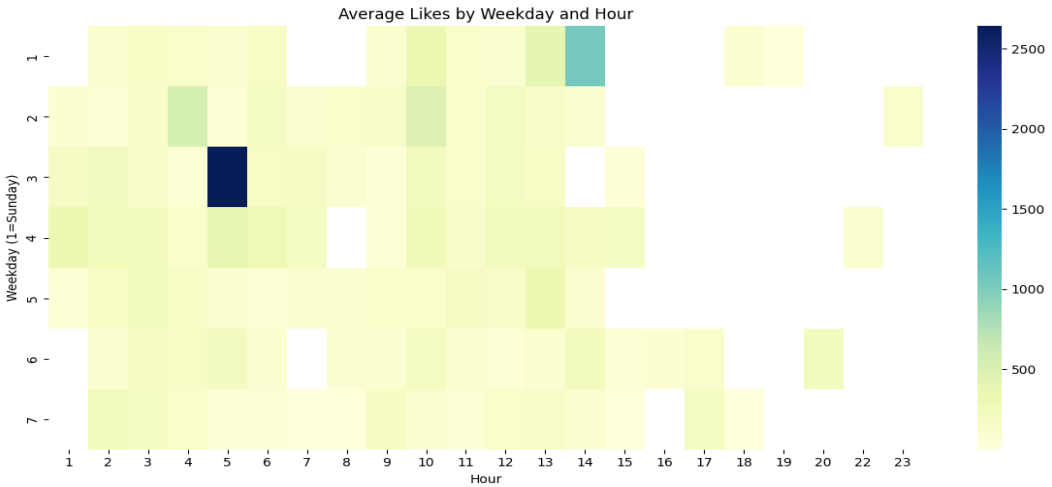

Figure 10. Average Likes by Weekday and by Hour

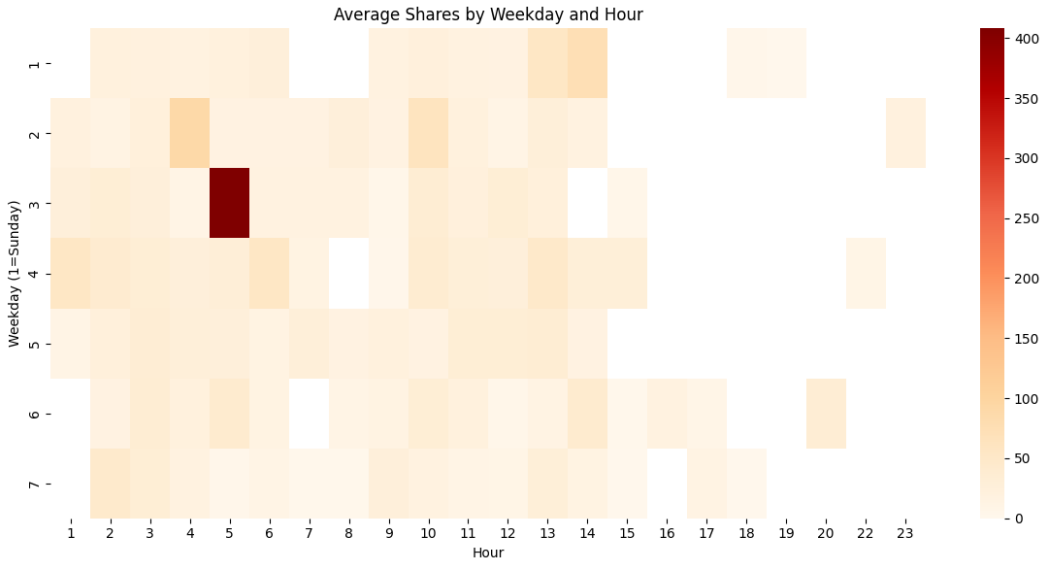

Figure 11. Average of Shares by the Weekday and Hours.

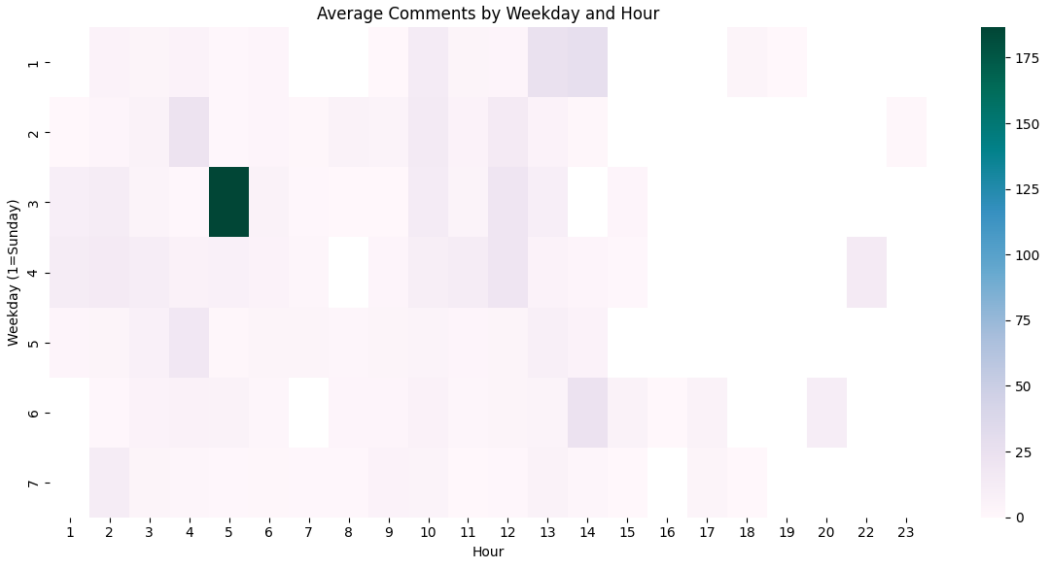

Figure 12. Average Comments by Day of week and by hour.

This indicates that there is a possible opportunity period that the journalists and media houses can take advantage and publish content that is reader attracting due to the low competition of users and the increased activities. Interaction is at its peak in the morning on mid-week days especially on Wednesday. The number of shares takes a similar direction to that of likes but at slightly high volatility.

4. Engagement Prediction Using Machine Learning Models

Commenting is less concentrated and nevertheless exhibits peaks in early hours during mid-week. In addition to the temporal trends, category-wise and hour-wise bars offer feasible information on the actionable content strategy. The average of likes and share of the Category 2 and 3 posts is always higher than that of Category 1, with a rather constant number of comments. In the same way, Figure 13

shows that the likes shoot through the roof at hour 5 and hour 14 again underlining the importance of timing in visibility.

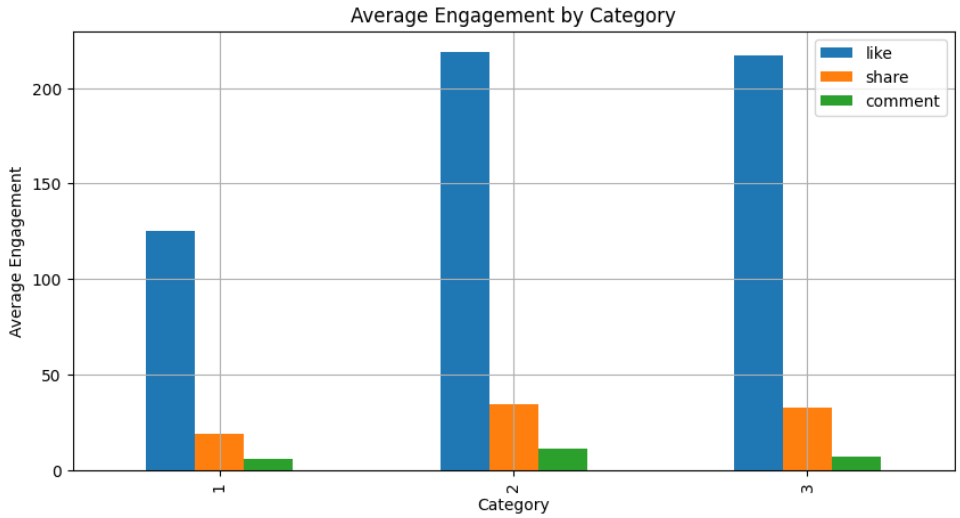

Figure 13. Mean Post Types Engagement.

The likes and shares in Category 2 and 3 are more as compared to Category 1 indicating more appealing content types.

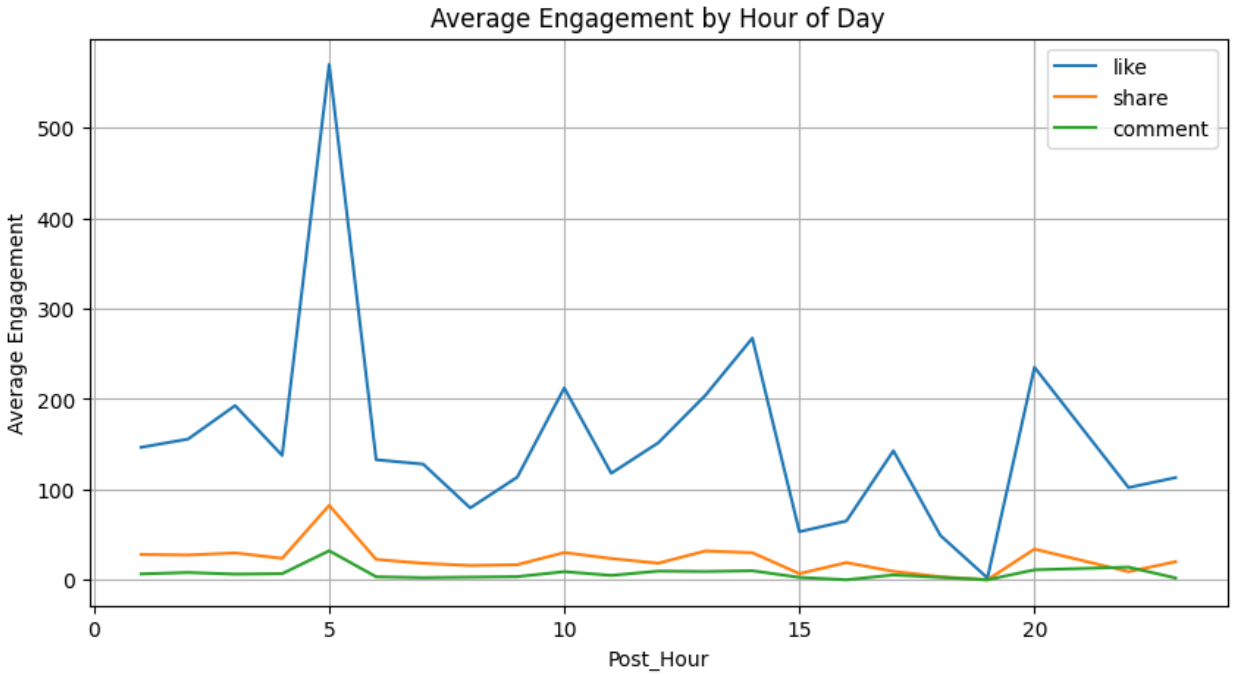

Figure 14. Average of Hourly Engagement.

Likes are highest at 5 AM and in the afternoon, to give one guidance on best time of scheduling. Finally, a boxplot distribution of Ring, based on the raw engagement data in Figure 15, demonstrates the existence of extreme outliers, in particular, in likes, proving the choice of log-transformation in the modeling process.

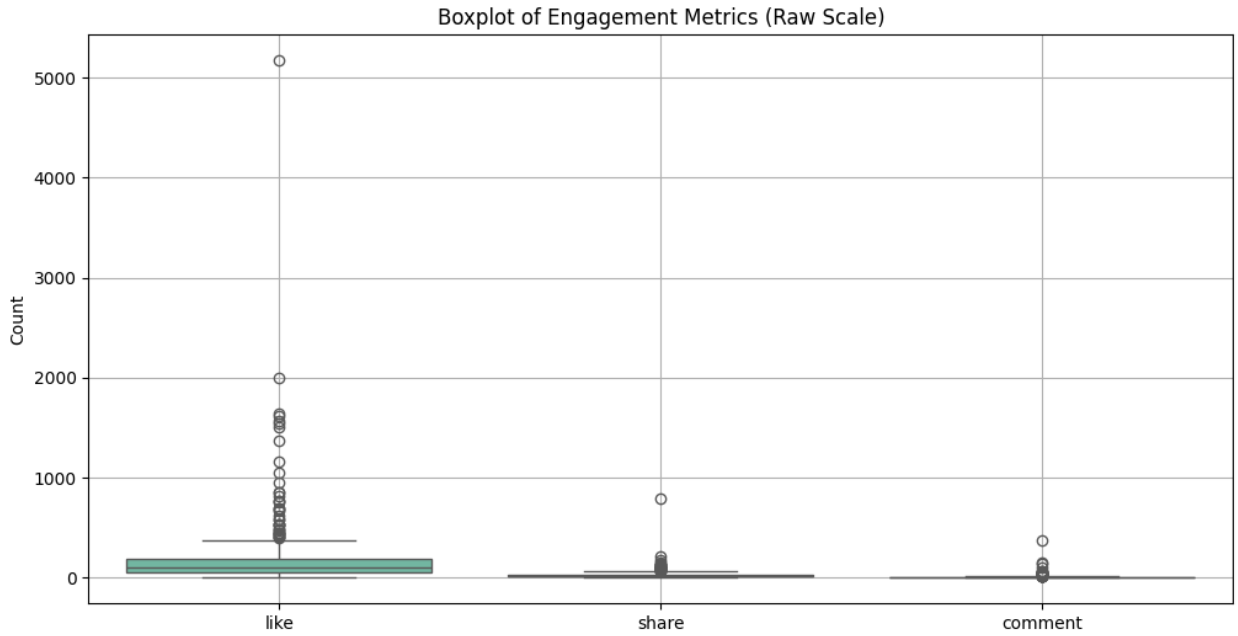

Figure 15. Box of the Engagement Metrics (Raw Scale).

There is high dispersion and outliers in the likes and shares and comments are more concentrated. The above three bar plots visualization shows the feature importance of predicting the metrics of Facebook engagement- the likes, shares, and comments through the XGboost regression model. Such visualizations can be used to perceive what features contribute most towards each type of engagement such as in figure 16.

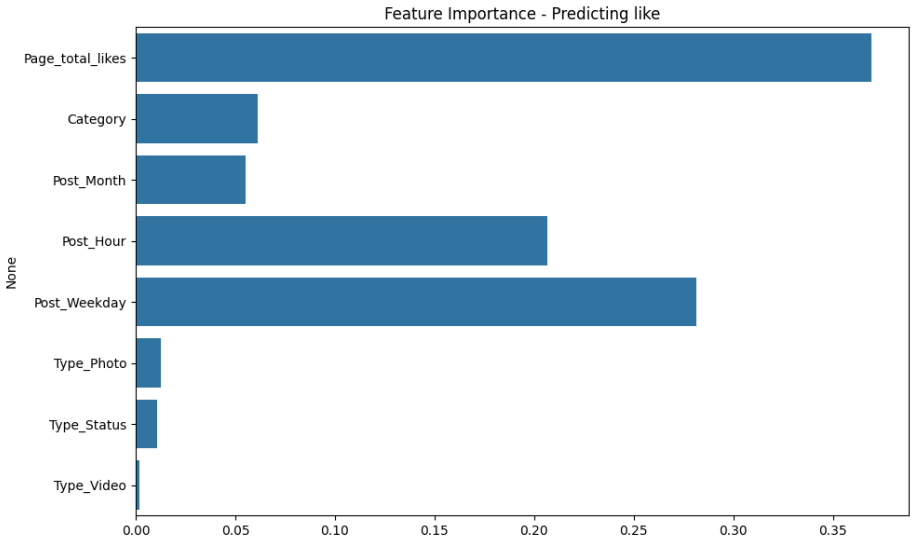

Figure 16: Feature Importance - Like Prediction

The most significant predictor of the number of likes a post gets is judged as Page_total_likes in this plot with an importance value of above 35%. There is then Post_weekday and Post_Hour which implies that the time of a post is also very important. By contrast, such variables as Type_Video and Type_Status have very few contributions, which mean that the post format does not play a significant role towards liking prediction as in figure 17.

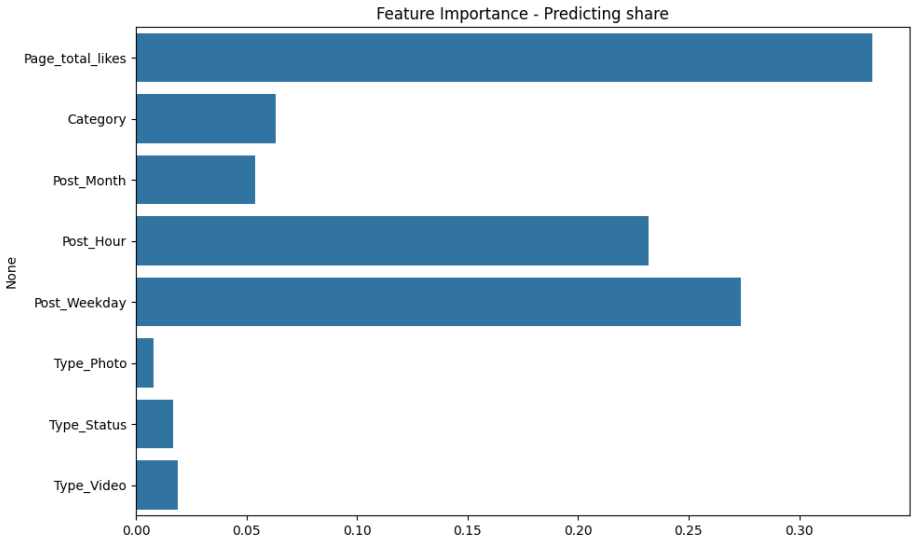

Figure 17: Feature Importance - Share Prediction

In the case of shares, Page_total_likes has been significant followed by Post_Weekday and Post_Hour that are close behind. This is to put it clear that it is the popularity of the page only but also the timing of the content that determines how much gets shared. Here, there is a little bit more impact of content type than likes but it is still rather insignificant as in figure 18.

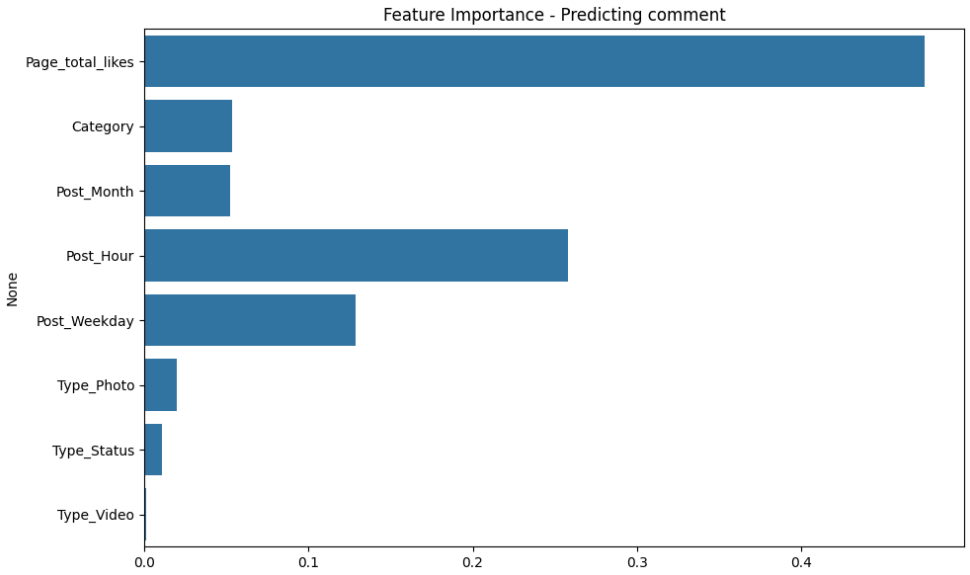

Figure 18: Feature Importance - Prediction Comment

Regarding comments, Page_total_likes also remains to be in the first place, but Post_Hour becomes even almost equal in importance. This means that commenting among users is more of a time-controlled aspect. Post_Weekday and content timing are again very influential, whereas categorical post types and months turn out to be weak predictors. In all three measures, Page_total_likes, Post_Weekday, and Post_Hour are always the top features, which further supports the conclusion that the popularity of a page and the time of posting are the main contributors to engagement.

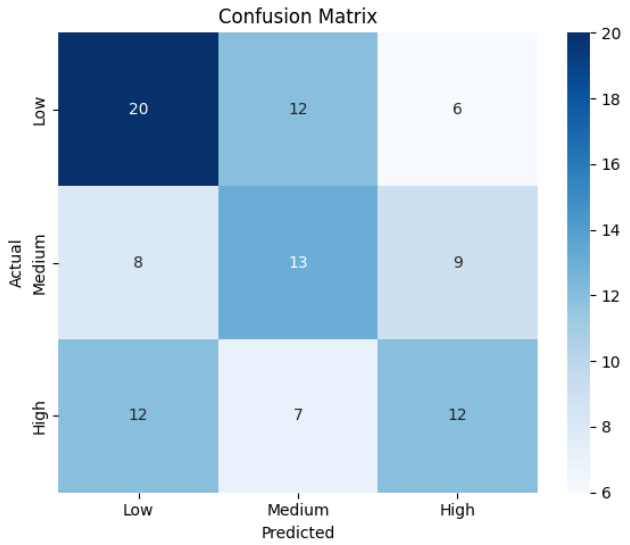

Figure 19: Confusion Matrix of Performance of the Model by Classes of Engagement

In this data set, post type and category of content provide little predictive ability pointing to a possible area of media teams to re-examine content type strategies. Such figure 19 has importance plots provide the transparency in the nature of machine learning modes as being black-boxes and also lead to meaningful decisions about social media optimization. The two confusion matrices provided provide graphical observations on how the model performs in the classification of engagement levels in Facebook post. Qualities of correct predictions are diagonal cells in figure 20. In the first confusion matrix (Figure 20) engagement is divided into Low, Medium, and High classes. The diagonal values of (20, 13, and 12) indicate the true positive results of each category. Nonetheless, there are also large off diagnals (e.g. 12 Mediums misclassified to Low, 9 Mediums misclassified to High) which suggests some moderate misclassification particularly between neighbouring classes. This implies that although the model is able to distinguish between broad range of levels of engagement, the small adjustments between Medium and other adjacent classes may prove difficult.

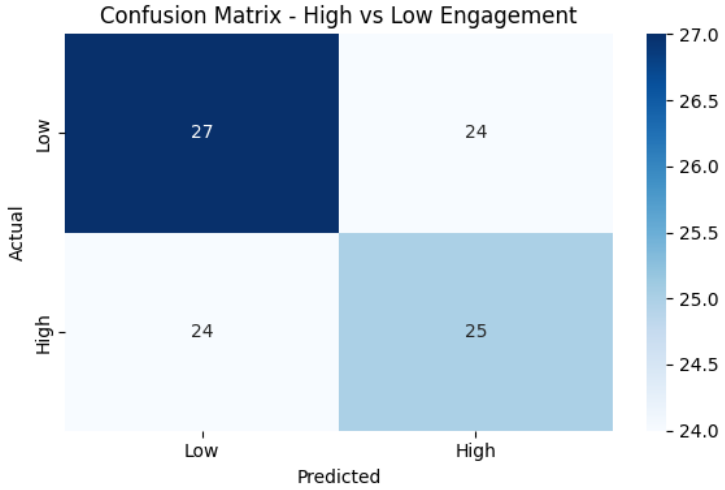

Figure 20: Confusion Matrix Of High vs Low Engagement

Equilibrated, yet cross-cut prediction zones gave them a hard time in figure 20. The binary problem of High vs Low engagement simplification found in the second confusion matrix (Figure 2) deceives the mind into thinking that the problem is simple. The performance is a bit more equal in this case with 27 and 25 of the true positives in Low-Low and High-High respectively. The misclassifications are also symmetric (24 each) indicating that though the binary approach mitigates the complexity, there is non-distinctive feature of high-performing and low-performing posts. Granted, it is perhaps a more viable schema to have such binary considerations to editorial rather than sophisticated content planning. The MAE values are obtained are shown in table 2 and plotted in Figure 21, and provide a further explanation of the relative accuracies of the models. XGBoost-GS recorded the smallest MAE in both log_like (0.41) and log_share (0.37), which is immediately surpassed by Gradient Boosting. This will mean a higher average proximity of the predicated to actuals. LightGBM once performed the best on the case of log_comment, with the lowest MAE (˜0.57), and Random Forest turned out to have the largest error (˜0.63). Gradient Boosting had lower MAE on every target; thus it is a good all-purpose method.

Table 2: Mean Absolute Error (MAE) of Likes Shares and Comments

Model	log_like	log_share	log_comment
XGBoost-GS	**0.4125**	0.3678	0.6175
LightGBM	0.4345	0.3726	**0.5680**
RandomForest	0.4407	0.4040	0.6295
GradientBoosting	0.4165	**0.3669**	0.5999

Figure 21: MAE Comparison across Models and Targets

The results outlined by the RMSE comparison chart (Figure 22) displays an evident pattern when considering the three target variables: log_like; log_share and log_comment. The XGBoost-GS model achieved the lowest RMSE two times than those of LightGBM and Random Forest, respectively. In addition, it had the lowest RMSE when it comes to log_like and log_share, holding an average RMSE of 0.59 and 0.48, respectively. Such findings have implications that the XGBoost-GS will be more accurate in predicting these engagement measures. The reasonable thing is that on log_comment LightGBM produced the lowest RMSE (˜0.75), which shows it was more stable on this noisier target. Random Forest had the largest RMSE in the majority of targets, especially in log_comment, which outlines its relative incapacity to overcome a wide range of variation in comment data.

Table 3: Root Mean Squared Error (RMSE) of Likes Shares and Comments

Model	log_like	log_share	log_comment
XGBoost-GS	**0.5901**	0.4875	0.7845
LightGBM	0.6074	**0.4840**	**0.7546**
RandomForest	0.6090	0.5296	0.8062
GradientBoosting	0.6143	0.4904	0.7673

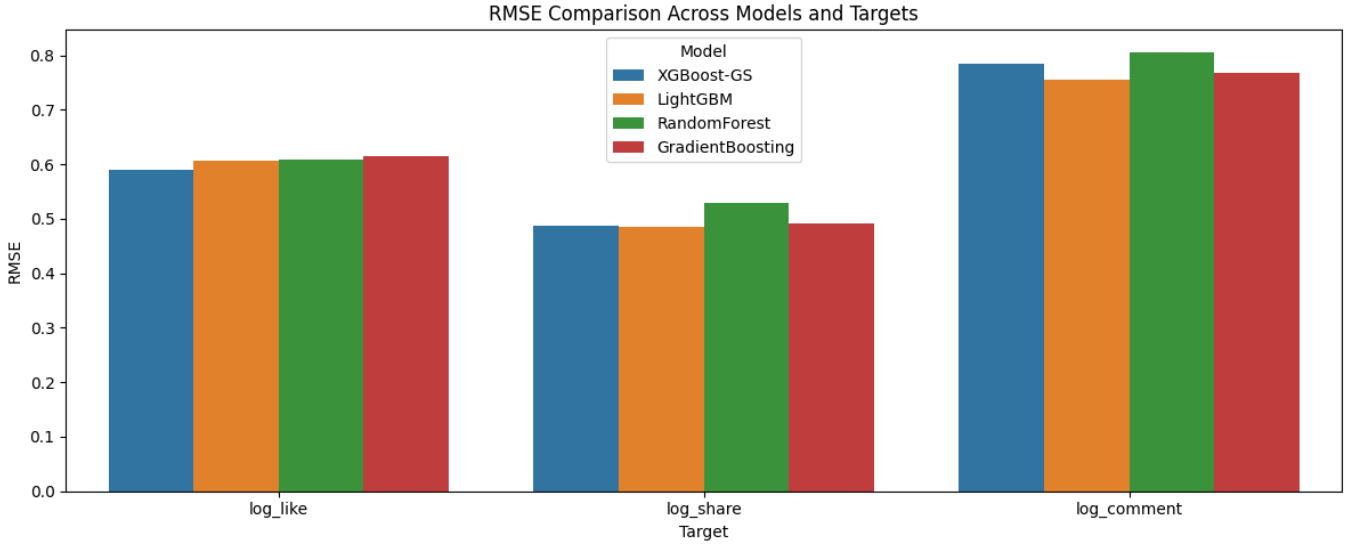

Figure 22: RMSE Comparison across Models and Targets

As it can be seen in Figure 22, the R2 Score comparisons support the previous results. In the case of log_like, XGBoost-GS returned highest R2 (~0.76) indicating that it explained a greater variance in likes. Gradient Boosting and LightGBM came next. The models that performed well on log_share included all but Random Forest which exceeded or stood close to 0.78, the LightGBM model performed best in terms of R square with about (0.785). The disparity is further accentuated in log_comment when once again LightGBM fronted the group (~0.49) and Random Forest brought up the rear (~0.42). These findings make the usage of boosting methods rather than bagging techniques more reliable in the given situation.

Table 3: R^2 Score of Likes Shares and Comments

Model	log_like	log_share	log_comment
XGBoost-GS	**0.7595**	0.7828	0.4521
LightGBM	0.7451	**0.7858**	**0.4931**
RandomForest	0.7438	0.7436	0.4213
GradientBoosting	0.7393	0.7801	0.4758

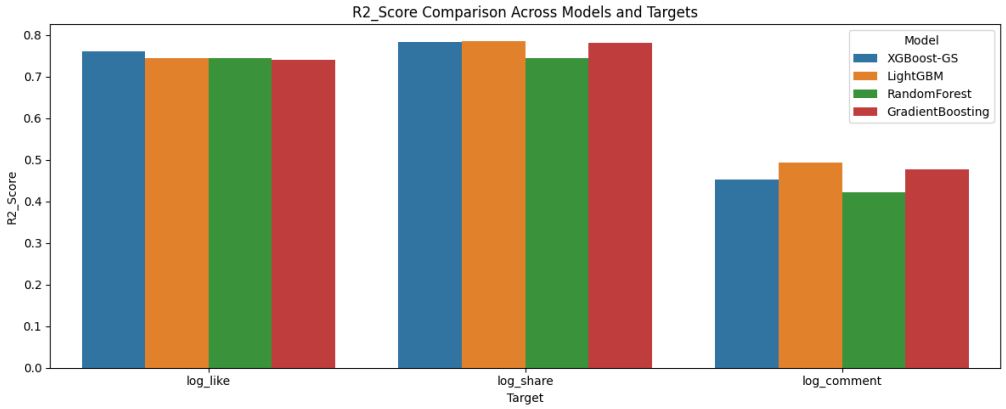

Figure 23. R2 Score Comparison across Models and Targets

The findings of this paper note that ensemble machine learning models are quite effective in mapping and predicting the later user engagement of Facebook posts, that is, the likes, shares, and comments. The model that has been consistently employed to give high results in predicting likes is XGBoost with GridSearchCV (XGBoost-GS) since it recorded the lowest MAE (0.4125), lowest RMSE (0.5901), and the highest ^{R22} ~ (0.7595). With respect to shares, LightGBM performed a slightly better result in RMSE (0.4840) and R2 (0.7858), but Gradient Boosting was the best in the forecast of MAE (0.3669), which leads to the conclusion that both these models are applicable in predicting this measure. Nevertheless, the prediction of comments was more complicated and unstable in all the models, and the overall best result was provided by the LightGBM model (MAE: 0.5680, R2 : 0.4931). Random Forest performed worse than boosting-based methods in every single target, and this was most probably because the random forests algorithm is biased towards specifics. The statistics obtained highlight the fact that predicting likes and shares can be done with good accuracy, but to enhance the models of predicting user comments it is possible to use certain subtler features like textual analysis or sentiment analysis to capture better results.

5. Conclusion

This paper showed a data-rich way of forecasting user engagement on Facebook with the help of ensemble machine learning models in journalism and mass communication. We used an organized data set of 500 posts and 19 variables and used the advanced statistical presentation of regression models; XGBoost with hyperparameter tuning, LightGBM, Random Forest, and Gradient Boosting to predict three significant measures of engagement; likes, shares, and comments. The models were evaluated by considering the MAE and RMSE, R2 values to determine their accuracy prediction and overall predictability. XGBoost-GS had the best overall results, especially in the prediction of log_like, with a MAE of 0.4125, RMSE of 0.5901, and a R2 Score of 0.7595, this meant that almost 76% of the variance in likes could be attributed to the model. In the case of log_share, LightGBM reported the highest R2 (0.7858) whereas Gradient Boosting had the lowest value of MAE (0.3669). By contrast, log_comment was by far the most difficult thing to predict, and LightGBM provided the best R2 (0.4931) because of the complex and volatile nature of user behavior in comments. Such results highlight the fact that likes and shares are more structurally predictable possibly due to quantifiable factors that include posting time, type and reach, whereas comments are influenced by subjective and contextual factors that include tone, controversial or topicality? this is because all these factors cannot be quantified and thus could not be captured at all in this dataset. The fact that R2 scores are relatively lower when making comment predictions is emphasized by the limitation of modeling engagement based on numerical and metadata characteristics only. Journalism and media plan wise, the outcomes have an operational value. Good engagement prediction can be used to plan the

editorial, frame content and target the audience by helping the newsroom and communication teams to fine tune how and when material should be delivered. Although the models used in this study demonstrate potential, future studies must combine the techniques of natural language processing (NLP), image classification, and sentiment analysis to bring into focus audience psychology behind comments and discussion of users of the comment section. Such collaborative practice would not only improve the quality of the model performance but facilitate media professionals with the powerful repeatable system of information-driven story and data-driven strategic approach towards digital engagement.

References

Ali, A., & Tajuddin, M. (2023). *Python for Journalism: Building Predictive Tools in the Newsroom*. Journal of Media Technology, 11(2), 55–70.

Alhuntushi, F., & Lugo-Ocando, J. (2020). Social media and the reshaping of journalism practices. *Digital Journalism*, 8(4), 456–472. https://doi.org/10.1080/21670811.2019.1697629

Carta, S., Corriga, A., Recupero, D. R., Satta, R., & Zola, F. (2020). A multi-layered and multimodal approach for user engagement prediction on Facebook. *Information Processing & Management*, 57(6), 102304. https://doi.org/10.1016/j.ipm.2020.102304

Elkalliny, M. (2021). The role of Facebook in health communication during COVID-19: Engagement metrics analysis. *Journal of Communication Inquiry*, 45(3), 281–299. https://doi.org/10.1177/0196859921998962

Ferrer-Conill, R., & Tandoc, E. C. Jr. (2021). The audience-oriented editor: Making sense of audience metrics in the newsroom. *Journalism Studies*, 22(6), 736–752. https://doi.org/10.1080/1461670X.2020.1853883

Jawley, T., & Fahmy, S. (2022). Reframing audience analytics in global journalism. *Journalism*, 23(4), 789–808. https://doi.org/10.1177/14648492110171740

Kalsnes, B. (2022). The power of platforms: Shaping news consumption and public discourse. *Digital Journalism*, 10(2), 261–278. https://doi.org/10.1080/21670811.2021.1990753

Tandoc, E. C., & Maitra, J. (2021). Analytics in the newsroom: A new editorial logic. *New Media & Society*, 23(5), 1267–1285. https://doi.org/10.1177/1461444820910406

Keco, D., Obucic, E., & Poturak, M**.** (2024). Improving the prediction of social media engagement in universities by utilizing feature selection in machine learning. *International Journal of Research in Business and Social Science*, 13(1), 372–380.

Arazzi, M., Cotogni, M., Nocera, A., & Virgili, L. (2023). Predicting Tweet engagement with Graph Neural Networks. *arXiv preprint*

Obucic, E., Poturak, M., & Keco, D. (2023). Predicting user engagement of Facebook post images in leading universities: A machine learning approach. *Revue d'Intelligence Artificielle*

Moro, S., Rita, P., & Vala, B. (2016). Facebook Metrics [Dataset]. UCI Machine Learning.

A. Shahzad, U. Farooq, A.I. Khattak, A.M. Durrani (2025). Predicting Online News Article Popularity across Social Platforms Using Machine Learning a Media Analytics Approach. *Journal of Media Horizons*, 6(2), 1096–1112. https://doi.org/10.5281/zenodo.15780381