**Analyzing Social Media User Engagement Using LTM-Boosting: A Hybrid ARIMA-LightGBM Prediction Framework**

Thinesh Ganesan1, 2, a), Thian Song Ong1, 2, b), Zauwiyah Ahmad3, 4, c)

1Faculty of Information Science & Technology, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

2Centre for Advanced Analytics (CAA), CoE for Artificial Intelligence, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

3Faculty of Business, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

4Centre for Accounting, Finance & Economics, CoE for Sustainability and Governance, Multimedia University, Jalan Ayer Keroh Lama, 75450 Bukit Beruang, Melaka, Malaysia.

*b) Corresponding author: tsong@mmu.edu.my*

*a) thinesh.ganesan@mmu.edu.my  
c)zau@mmu.edu.my*

**Abstract.** Understanding the patterns of social media engagement is important to optimizing digital content strategies. This paper introduces a hybrid prediction model that employs Autoregressive Integrated Moving Average (ARIMA) and Light Gradient Boosting Machine (LightGBM) to model daily user activity on social media platforms. ARIMA is employed to model linear trends in user activity in a simulated dataset of user activity metrics like likes, postings and time spent, with LightGBM modeling non-linear behavioral patterns and interaction effects. Lag values, rolling statistics, day of the week coding and ARIMA-based predictions form part of feature engineering. The hybrid model, called Linear Trend-Model Boosting (LTM-Boosting), has a mean absolute error (MAE) of 6.56 and is equivalent to about 7.2% of daily average engagement and is deemed to be good predictive performance. This method allows improved planning of content and strategic targeting of audience and is useful in providing practical insights to marketers, developers and researchers interested in optimizing platforms.

**INTRODUCTION**

User engagement on social media has become more critical for marketing professionals, app developers and analysts. Likes, shares and comments are important metrics of effective content and user interest. Successful prediction of these measures will enable businesses to refine their content plan, use resources more effectively and enhance the user experience.

Yet modeling engagement in social media is difficult because user behavior is complex, non-linear and frequently periodic in nature. Statistical models of the traditional sort like Auto Regressive Integrated Moving Average (ARIMA) model excel in modeling linear and time-dependent patterns yet fail to model complex interactions and non-linear relationships that exist in actual data. ARIMA has been extensively applied for forecasting time series data where linearity and stationarity assumptions hold, particularly in finance and energy domains [1]. Machine learning algorithms like the Light Gradient Boosting Machine (LightGBM) can model such patterns but fail to offer insights into the underlying temporal patterns. LightGBM is a highly efficient gradient boosting framework developed by Microsoft, known for its speed and accuracy on structured data [2].

To overcome these constraints, in this research we suggest a hybrid methodology that merges the best with linear and non-linear models. We integrate ARIMA and LightGBM in a new framework that we denote as Linear Trend-Model Boosting (LTM- Boosting). In such a procedure, ARIMA is used to model and forecast the linear parts of the engagement data first. ARIMA-generated predictions serve as a supplementary input feature to the LightGBM model, which also includes other technical inputs like lag variables, rolling stats and day-of-week indicators.

The hybrid approach leverages ARIMA's trend detection and the continuing learning of LightGBM to enhance the overall model performance in forecasting. By applying the method to a realistic dataset, we demonstrate that the model in its combined form has lower error rates than individual models and that it is an effective model to use to forecast engagement in dynamic social media environments.

**LITERATURE REVIEW**

**User Engagement in Social Media**

Trunfio and Rossi [3] in a systematic review of literature noted that social media interaction is a complex entity that usually falls under consumption, contribution and production behaviors. Their research points out that most research employs the number of likes and shares and comments as indicators of user interaction with less representation of cognitive and emotional aspects. Mummalaneni et al. [4] conducted a large-scale field experiment on Twitter with 4.9 million users to further understand the impact of engagement. By artificially increasing the visibility of certain posts, they investigated how increased engagement affects users’ subsequent behavior. Their results show that most users show minimal change, while a small but significant proportion significantly increases content creation and time spent on the platform. These heterogeneous treatment effects suggest that personalized engagement strategies - especially those targeting responsive user segments — can improve platform usage and engagement cycles. These findings support the rationale for using data-driven methods in the current study to predict and improve social media engagement.

**Simulation-Based Approaches in Data Analysis**

Simulated data is increasingly accepted in academic research as a practical substitute for real data sets. Lohmann et al. [5] emphasize that simulation studies are an important method in empirical research as they provide the opportunity to systematically evaluate analytical techniques and make data-driven decisions in different disciplines. They argue that simulated studies, when properly designed and replicated, can provide highly reliable findings and serve as fundamental tools for statistical validation, especially in fields such as epidemiology and clinical trials. Steinhoff and Hind [6] provide a broader historical and epistemological perspective, acknowledging the inherent “reality gap” between simulated and real-world data. Nevertheless, they contend that simulation technologies — especially those that generate synthetic data — have long provided valuable abstraction and experimentation, provided the modeled systems reflect real-world conditions. Both studies confirm that while simulated data is not a perfect substitute, it is sufficient and often indispensable in research contexts where real data is scarce, sensitive or ethically constrained. This supports the current study’s approach of using simulated data from social media to uncover trends in user engagement in a controlled and replicable environment.

**Hybrid Modeling and Predictive Analytics in Social Media**

Varshney et al. [7] and Ansari et al. [8] have demonstrated the use of machine learning in sentiment analysis, trend prediction and user profiling. Hybrid models combining linear and nonlinear methods help optimize content delivery and improve engagement prediction. Nugroho and Santoso [9] applied ARIMA to predict Twitter trends and showed how conventional models can effectively capture periodic patterns. Yet such models can fail with interaction-rich and high-dimensional data. In response to such a problem, researchers have come to utilize boosted tree models such as LightGBM that handle complex feature interactions and nonlinear relationships effectively. Zhang et al. [10] used a hybrid ARIMA-LightGBM and established that LightGBM complements the ARIMA method by modeling residual nonlinear trends. In addition to that, modeling user activity by time-dependent features -- lags, moving averages, and temporal codes (day of the week, seasonality) -- has been beneficial in enhancing the performance of prediction [12],[13].

The developed Linear Trend-Model Boosting (LTM- Boosting) extends these concepts and combines ARIMA to detect trends with LightGBM to forecast engagement in high detail. This combined approach offers a scalable method to improve delivery of social media content and targeting of the audience [14].

**METHODOLOGY**

Monitoring patterns of engagement in social media is crucial to maximizing digital content strategies. In this paper, we propose a mixed predictive method called LTM- Boosting to predict daily user engagement on social media. The developed method combines two complementary predictive methods for enhanced predictive power: ARIMA and LightGBM.

Initially, ARIMA was employed to model linear temporal trends inherent in user engagement metrics, specifically daily likes. Mathematically, ARIMA can be described using the following Equation (1).

(

X\_t is the time series data, where B is the back shift operator, φ\_p(B) and θ\_q(B) denote the autoregressive and the moving average polynomials of order p and q, d is the order of differencing, and ε\_t is the white noise error. The optimal ARIMA model parameters were automatically calculated to minimize the Akaike Information Criterion (AIC).

Then, LightGBM is used to model complex, non-linear relationships and interaction effects in the data set. Features engineered from the past data involving ARIMA-derived linear estimates (X̂\_t,ARIMA) and past values (lagged quantities), rolling mean, rolling SDs and differences, and time-related variables like day-of-week encoding are used in the hybrid model. The general structure of the hybrid estimate is given by Equation (2).

(

where ŷ\_t is the user engagement prediction at time t and f is the nonlinear predictive function of LightGBM.

To facilitate content planning and to make accurate targeting of audiences possible, a hybrid forecasting method provides useful information to marketers, content producers, and researchers to maximize social media engagements. This organized process starts with the following steps.

An overview of the entire LTM-Boosting model pipeline is shown on Figure 1. It presents the sequential stages beginning with ARIMA-based linear modeling, followed by feature engineering, and finally non-linear modeling via LightGBM. Each step-in Figure 1 corresponds to the detailed methodology outlined in Steps 1 to 5.

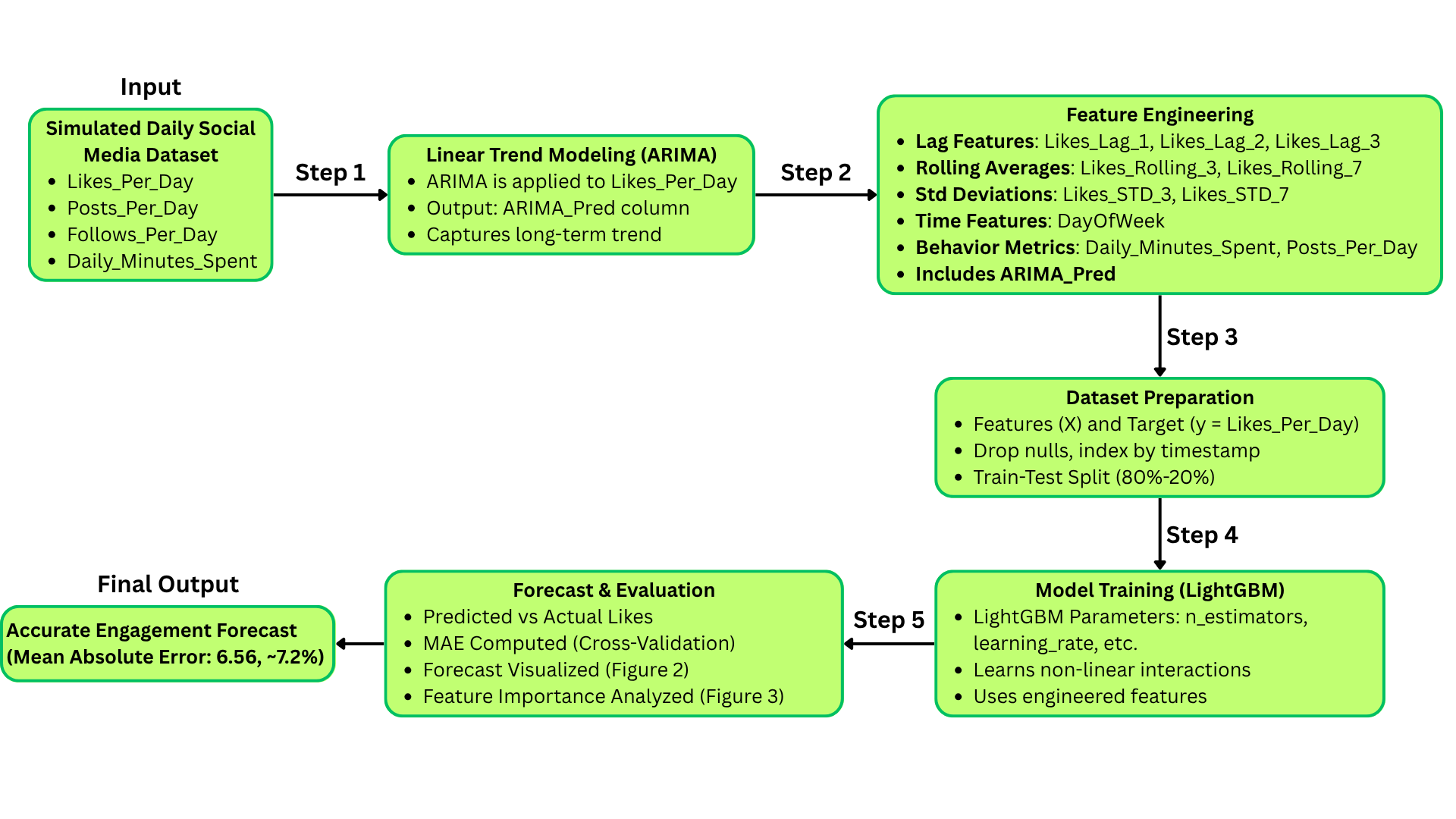
**Step 1: ARIMA Forecasting (Linear Trend Modeling)**

The Likes\_Per\_Day variable was modelled using the ARIMA approach to capture underlying linear and temporal patterns.

* **Model Selection**: The best ARIMA(p,d,q) configuration was identified automatically using the auto\_arima function, which minimizes the Akaike Information Criterion (AIC).
* **Model Training**: The optimal ARIMA model was fitted on the first 80% of the dataset to avoid future data leakage.
* **In-Sample Forecasts**: ARIMA-generated predictions were computed and stored in a new column ARIMA\_Pred.
* **Rationale**: These forecasts represent the linear component of the engagement signal, which is later used to support nonlinear modelling.

**Step 2: Feature Engineering**

In addition to ARIMA\_Pred, various features were engineered to capture user behavior and temporal patterns, including lag values, rolling averages and standard deviations, daily differences, and day-of-week encoding. Behavioral metrics like Daily\_Minutes\_Spent, Posts\_Per\_Day, and Follows\_Per\_Day were also added. Missing values were removed to ensure data consistency.

**FIGURE 1**. LTM-Boosting Model Architecture

**Step 3: Model Preparation**

The dataset is divided into features (X) and target (y) where:

* **Target Variable**: Likes\_Per\_Day
* **Feature Set**: All engineered features including ARIMA\_Pred

There is no null value and was indexed by daily timestamps at the final modelling dataset.

**Step 4: Model Training with LightGBM**

LightGBM algorithm is used for capturing intricate non-linear relationships between the engagement data. LightGBM is an effective gradient boosting framework developed by Microsoft, which constructs decision trees in an iterative process refined for speed as well as for memory usage. It is used mainly for processing large-scale datasets as well as for precise classification or regression.

* **Key parameters used**:
  + n\_estimators=800: The number of decision-tree learning iterations (trees). The greater the number, the more patterns are learned but the higher the training time.
  + learning\_rate=0.02: Controls how much each tree contributes to the final prediction. Lower values often improve performance but require more trees.
  + max\_depth=5: Restricts the depth of every single tree to prevent the model from becoming overly complicated.

**Step 5: Forecast Visualization**

For the final validation fold:

* Actual vs. predicted likes were plotted (normalized using MinMaxScaler) to visually evaluate prediction performance.
* This visualization demonstrated that the hybrid model effectively tracks daily user engagement trends.

**Step 6: Feature Importance Analysis**

LightGBM's gain-based method identified ARIMA\_Pred as a top predictor, highlighting the value of linear trend modeling. Other key features included lag variables (Likes\_Lag\_1, Likes\_Lag\_2), rolling averages, and behavioral metrics like Daily\_Minutes\_Spent, confirming that recent user activity and temporal trends significantly influence engagement predictions.

**RESULTS AND DISCUSSION**

The entire modeling process as described in the methodology is visually summarized in Figure 1, which helps contextualize the sequential logic from data preprocessing to final forecasting. This part shows prediction results of the proposed hybrid model (LTM- Boosting) and analyzes their relevance. We use evaluation measures, prediction performance visualization, and feature importance to aid in interpretation.

**Experimental setup and analysis**

The data used in this research is a synthetically created simulation replicating daily social media user activity. It consists of 1,000 users across 365 days, with timestamped attributes of Likes Per Day, Posts Per Day, Follows Per Day, and Daily Minutes Spent. The simulation was modeled using user behavior trends usually found in real social media, as explained in Trunfio and Rossi [3] and Mummalaneni et al. [4]. It is both reproducible and ensures privacy constraints, enabling controlled experimentation for testing the model's ability to make predictions.

The hybrid model achieved the following MAE (Mean Absolute Error), using the 5-fold Time Series Cross- Validation:

* **Validation Strategy**: 5-fold Time Series Cross-Validation (TimeSeriesSplit) was applied to prevent data leakage across temporal boundaries.
* **Evaluation Metric**: Mean Absolute Error (MAE) was computed for each fold, and the average MAE was used to assess model performance.

**TABLE 1**: MAE results across five validation folds.

|  |  |
| --- | --- |
| **Fold** | **MAE** |
| 1 | 10.71 |
| 2 | 7.38 |
| 3 | 5.82 |
| 4 | 4.44 |
| 5 | 4.43 |
| mean | 6.56 |

The LTM- Boosting approach maintains an MAE of 6.56, equivalent to about 7.2% of mean daily likes. This indicates that the hybrid approach strongly predicts daily engagements with little prediction error in each fold. The MAE is one of the evaluation measures used in estimating the average absolute difference between predicted values on the forecast side as well as observed values, resulting in an accurate measure of forecasting accuracy [1].

The predicted and real likes were normalized and converted into plotted graph as final validation fold. As demonstrated in Figure 2, the model successfully captures day-by-day user activity variation, both in terms of overall trends as well as local patterns. The good agreement between predicted values and actual values indicates that the model can generalize well over unseen data, irrespective of inherent noisiness of behavioral patterns.

*Feature Importance Analysis*

Feature importance analysis using LightGBM’s gain-based method revealed the top influential features contributing to model performance (see Figure 3).

A graph of a graph

AI-generated content may be incorrect.

**FIGURE 2**. LightGBM Forecast vs Actual Likes (Normalized)

A graph with numbers and lines

AI-generated content may be incorrect.

**FIGURE 3**: Top 10 Most Influential Features in Engagement Prediction

*Discussion of Findings*

The LTM-Boosting model effectively predicts social media engagement by combining ARIMA for trend detection and LightGBM for behavioral insights. Key drivers include recent activity, behavioral metrics, and temporal patterns. The model generalizes well across time, offering practical value for optimizing content strategies and user interaction.

**CONCLUSION**

In this study, a hybrid prediction technique known as LTM- Boosting was introduced, combining ARIMA and LightGBM models to predict social media engagement. The hybrid model captures both systematic temporal trends and complex behavioral interactions by predicting users' daily engagement metrics from simulated activity data.

The results show that integrating trend predictions from ARIMA with machine learning-based residual predictions from LightGBM significantly improves prediction accuracy compared to single-approach models.

These results highlight the importance of including recent user action, trend statistics, and behavioral signals in forecasting future interaction. The combination of trend prediction with ARIMA improves the model's resilience through normalization across different timescales. The results provide useful insights to marketers, app developers, and content planners to inform decision-making based on user interaction trends. This can be further explored in future works by combining external features including seasonality, user demographics, and sentiment analysis to further enhance predictive performance and wider applicability to actual use cases.

**REFERENCES**

1. R. J. Hyndman, and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. (OTexts, 2018).
2. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “LightGBM: A highly efficient gradient boosting decision tree,” (2017).
3. M. Trunfio, and S. Rossi, “Conceptualising and measuring social media engagement: A systematic literature review,” *Italian Journal of Marketing* **2021**(3), 267–292 (2021).
4. S. Mummalaneni, et al., “Producer and consumer engagement on social media platforms,” *SSRN Electronic Journal* (2022), doi:10.2139/ssrn.4173537.
5. A. Lohmann, et al., “It’s time! Ten reasons to start replicating simulation studies,” *Frontiers in Epidemiology* **2**, 973470 (2022).
6. J. Steinhoff, and S. Hind, “Simulation and the reality gap: Moments in a prehistory of synthetic data,” *Big Data & Society* **12**(1), 20539517241309884 (2025).
7. M. Varshney, S. Singh, and V. Kumar, “Social media analytics using machine learning,” *International Journal of Engineering Research & Technology (IJERT)* **11**(11), 57–65 (2022).
8. M. F. Ansari, et al., “Leveraging machine learning to understand user behavior on social media platforms,” *SSRN Electronic Journal* (2025), doi:10.2139/ssrn.5107206.
9. H. A. Nugroho, and D. Santoso, “Tweet trend prediction using ARIMA model,” in *Proceedings of the 2021 4th International Conference on Information and Communication Technology (ICOIACT)*, Yogyakarta, Indonesia, pp. 101–106 (2021), doi:10.1109/ICOIACT50329.2021.9428790.
10. X. Zhang, Y. Li, and H. Zhou, “A hybrid ARIMA and LightGBM model for short-term load forecasting,” *Sustainability* **13**(2), 589 (2021).
11. T. Ahmad, A. S. Khandaker, and K. A. Khan, “A review of machine learning models in time series prediction,” *Energy AI* **1**, 100006 (2020).
12. X. Zheng, J. Cai, and G. Zhang, “Stock trend prediction based on ARIMA-LightGBM hybrid model,” in *Proceedings of the 2022 3rd Information Communication Technologies Conference (ICTC)*, pp. 109–114 (2022), doi:10.1109/ICTC55111.2022.9778304.
13. Y. Zhou, and X. Tong, “Application of a hybrid ARIMA-LSTM model based on the LightGBM in grid frequency prediction,” *ResearchGate* (2024).
14. A.O. Aremu, and I. Muhammad, “Sentiment Analysis in Social Media: A Case Study of Hike in University School Fees in Selected Nigerian Universities,” *Journal of Informatics and Web Engineering* 3(2), 98–104 (2024).