Anusandhanvallari
Vol 2024, No.1
April 2024
ISSN 2229-3388

"AI-Powered Modeling for Behavioral Forecasting: A Hybrid Approach to Social Media User Behavior Analysis"

[1]Dr. Kiran Chaudhary, [2]Aakash Punit

[1]Associate Professor, Department of Commerce, Shivaji College, University of Delhi

[2]Assistant Professor, Department of Commerce, Kirori Mal College, University of Delhi

Abstract

The boom of social Networking sites generates tons of data also known as rich, massive and social diffused data for user preference, sentiment and behavior capturing. The hybrid approach proposed in this proposal aims to improve social media user behavior prediction ability. Random Forest realize feature importance rank and dimension reduction whereas Long Short-Term Memory (LSTM) model uses the temporal behavior. The web data work as a fuel of the processing pipeline with fetching from Twitter, and Instagram "Realtime" up to submodule for NLP preprocessing of data and training of supervised models. Based on comprehensive performance metrics, the proposed ensemble model based on Random Forest and Long Short-Term Memory networks (RF-LSTM) yields significant superiority to feature-based single classifiers with respect to accuracy, F1-Score and RMSE for trends in user engagement expression prediction as well as that of sentiment polarity and interaction patterns. Our findings highlight how the feature selection step using a Random Forest, can improve the performance of LSTM on high dimensional noisy datasets (such as social media data). The hybrid framework further enhances generalization, interpretability and thus becomes suitable for deployment in dynamic digital environments. Besides, certain scenarios for human anticipative perturbation detection in SNS targeted advertisement, policy monitoring and psychological prevention could be on the top of the design compositions to predict if a user will elaborate any reaction regarding their behavior via digital well-being service. Our approach provides an end-to-end solution that is widely validated and applicable across all behavioral domains, unlike this line of work.

Keywords: Social Media Behavior Prediction, Random Forest–LSTM Hybrid Model, Behavioral Forecasting, Artificial Intelligence in Social Networks, User Engagement Analysis

1. Introduction

1.1 Background and Motivation

The emergence of SNS online Social Media platforms, (e.g., Twitter, Instagram or Reddit) has reshaped digital communication and opened up new avenues for behavior data analysis. On the other hand, these channels are a driving force of content sharing and retrieval of user intent perception sentiment and engagement behavior (Hochreiter, S., & Schmidhuber, J.1997). In 2023 alone, over 4.9B users interact upon social media across a global scale (yet to come), and the work of analyzing user behavior and predicting actions has been important in numerous applications including digital marketing, public health surveillance, political prediction and disaster reduction (Jiang, G., & Cottrell, G,2017).

Artificial Intelligence (AI), in particular with the advent of machine learning (ML) and deep learning, has thus become a powerful toolbox to realize knowledge discovery out of such unstructured and constantly developing data. Recently, the ability to handle time series data has been investigated through application of niche techniques

Random Forest (RF) and sequence based deep learning models Long Short-Term Memory (LSTM) network (Liaw, A., & Wiener, M,2002). This hybridization combines the ability of RF to learn high-dimensional data with the capability of LSTM to capture complex temporal patterns, proposing more accurate behavior predictions for users.

1.2 Research Problem and Objectives

Related work even though, need for the study on social media analytics is emerging interest but large numbers of models have been found to suffer with noisy data (Chen, T., & Guestrin, C,2016), non-linear data and temporal data like user interaction data. Existing approaches either being focused on ranking important features (to make the predictions) or capturing time-series dependency, enable non-deceptive predictive performance assessment but lack interpretability.

- To overcome the mentioned limitations, this paper proposes an AI-augmented hybrid model based on Random Forest and LSTM for predicting user engagement patterns in social media networks. The detailed aims are:
- Most importantly, to develop a hybrid AI BE model for predictive analysis of behaviors that is end-to-end interpretable and scalable
- To analyze the prediction performance of RF−LSTM ensemble on live social media data.
- Articulate the applicability of this model in actual fields such as digital marketing and user engagement optimization

1.3 Significance of the Study

The innovation of this work is that it is a full-framework mixture of the hybrid formulation in the static feature evaluation and dynamic sequence learning. Unlike theoretical work or simulation-based work, this paper provides a usable model pipeline and validate the model on real-world datasets (GBiau, G., & Scornet, E,2016). The results obviously have direct relevance for social media strategists, content personalization engine and public policy platforms, all looking towards a proactive reaction to changes in user behavior.

It is also going to guide hybrid AI applications in behavior-centric fields, thus a methodological contribution for behavior prediction and temporal user models (Qi, Y., Chen, Q., & Li, Y,2023).

1.4 Structure of the Paper

The rest of the paper is structured as follows:

- **Section 2** reviews related literature and highlights research gaps in behavioral forecasting using AI.
- **Section 3** describes the methodology, including data sources, preprocessing, Random Forest feature extraction, and LSTM model design.
- **Section 4** outlines the technical implementation of the hybrid model and deployment setup.
- **Section 5** presents performance results, visualizations, and comparisons with baseline models.
- **Section 6** offers a critical discussion of findings and their alignment with previous research.

- **Section 7** explores practical applications and implications for real-world deployment.

- **Section 8** concludes with key takeaways and directions for future research.

2. Literature Review

2.1 Background of Online Predictions of Behavior

Even predicting user behaviors in the digital world has roots in behavioral sciences and more recently computational modeling when social media platforms became darlings of the information ecology. These types of behavioral predictions on the platforms utilize user interactions such as thumbs, sharing activity (having to deal with other like comments) and publishing rate. Earlier models were largely scaling on a rule-based or regression-based algorithms, which failed to represent the dynamic nature of interactions including online active users (Probst, P., Wright, 2019). But with the advent of data in its unstructured data form during various platforms, the need was booming for scalable/adaptive models which can also evolve based on such huge volume of data.

2.2 AI and ML for Social Media Behavior Analysis

The continuous increase of Artificial Intelligence (AI) and Machine Learning (ML), data collection has become more powerful in tracking and predicting model users' online social media activity behavior. Models like decision trees, SVM, k-NN etc. formed the basis of predictive tasks in sentiment analysis, trend detection, content recommendation etc., (Mohammed, A., & Mohammed, A.,2024). However, the classic techniques would have bad performance to meet high-dimensional feature space and also ignored temporal dependencies between user interactions.

Recent evidence has shown that the best of social media analysis comes from a hybrid blend of AI methods. For instance, Li and Zhao (2022) proposed the design of an LSTM network for predicting time-dependent engagement, whereas (Patel and Joshi) proposed a combination between Random Forest and deep learning in behavior classification task, where hybrid algorithm gets improved accuracy along with better interpretability.

2.3 Application of Random Forest for Feature Selection

Random Forest (RF) has been commonly used in both classification and feature selection tasks with high dimensional data, i.e., social media data. RF can find the most important predictor variable by measuring the significance of variables across multiple runs because of the ensemble and decision tree forms (Jiang, G., & Cottrell, G.,2017).

Wang and Liu (2021), used RF on the selected feature to improve sentiment analysis by extracting significant feature combination from available corpus in textual data space. Kim and Choi (2020) demonstrated that the RF feature selection approach results in improved performance of deep learning user activity prediction classifiers. In our paper, we adopted RF as a procedure pre-processing module to remove noise and sparsity data, and select feature behaviors that were able to enter the time-series model.

2.4 LSTM for User Temporal Behaviors Modeling

Long short-term memory (LSTM) networks become a powerful model for learning temporal patterns from sequential data like user activities being affected by various time phases. In contrast to history RNNs, LSTMs alleviate the vanishing gradient problem and hence have long memory capabilities (which is sought in examining user-engagement behaviors per post, sessions s or time-windows) (Darabi, H., & Chen, S,2018).

Using time sequence of comments, (Zhang and Chen ,2024) applied the LSTM model in online community data, uncovering that it surpassed models without temporal prediction based on comment frequency and stay time. (Wang and Zhang ,2023) uses LSTM layers along with attention mechanisms to predict future interactions of a given user at high temporal resolution. These only works focus on user whose data is complex and varies with time so they rely on using LSTM to handle that complexity.

2.5 Research Gaps Identified

Although RF and LSTM have drastically improved behavior prediction on their own, a systematic study of both has not been well studied. These models are either static model based which loses the temporal pattern (Černocký, J., & Khudanpur, S,2010) or simply deep network based but can easily be overfitting and lack interpret ability in the absence of prior attentive feature filtering (Tavakoli, N., & Namin, A. S,2019).

In addition, also due to the lack of end-to-end pipelines that close the gap between hybrid models and forecasting methods similar to those that people use in systems with real-time requirements, these methods are poorly justified. In addition to this (Gundecha, P., & Liu, H,2012) also pointed out that very few studies are conducted for practical case study analysis comparing multiple hybrid models against real datasets with respect to prediction and operational feasibility.

This paper addresses these gaps and presents a concrete implementation of Random Forest–LSTM framework that has combined feature importance with temporal learning. The model is interpretable, scalable and ready for immediate use in domains such as marketing automation, behavioral targeting and digital public policy.

3. Methodology

3.1 Research Design and Methodology

The method utilizes a hybrid supervised machine learning model combining Random Forest (RF) feature selection with Long Short-Term Memory (LSTM) networks for time-series forecasting. The aim is to maintain both the static\\ behavior characteristics, as well as the dynamic properties of social user interactions.

The training of the model smashes full chain from data gathering, preprocessing, feature engineering, sequence modelling until performance evaluation Using common social media data, the scholarly research focuses on application of both frameworks for predictive behavior with algorithm trained and tested methods against a variety of traditional approaches.

To verify all components of the framework from practical perspective experimental method with settings on facility is followed. This allows for one the ability to assess and compare performance but also patterns of behaviors across time windows.

3.2 Primary Data Sources and Evaluation Criteria

It was mined from various publicly available social data (as they have high interaction rates, including different types of user engagement; commentary vs intellectual debate etc.) The below data was scraped via official APIs:

Twitter: Disciplining of profit companies we've never voted for.

Instagram: Hashtag extraction with base=post - extract metadata (views, comments count, timestamp).

For January to March 2024 only the English language articles were taken. To preserve data quality, users with fewer than 50 interactions throughout the observational period were removed.

Table 1 summarizes the volume, source, and filtering criteria of the dataset:

Table 1: Summary of Collected Social Media Data

Platform	Data Fields Collected	Time Frame	Total Records	Filtering Criteria
Twitter	Tweet text, retweets, likes, replies, timestamp	Jan–Mar 2024	180,000	English only, >50 retweets or replies
Instagram	Caption, hashtags, likes, comments, timestamp	Jan–Mar 2024	95,000	Public posts, hashtags >10k usage

3.3 Data Preprocessing and Cleaning

To transform this raw text into a usable dataset, we utilized multiple Natural Language Processing (NLP) techniques:

- Text was standardized using tokenization and lemmatization.
- It made OLDER stop words, special characters and hyperlinks disappear.
- We calculated sentiment scores using the VADER sentiment analysis tool.
- Timestamps were actively adopted to keep track of time-ordered sequences of activities.

For non-textual data:

- We used min-max normalization for numerical variables (such as likes and comments).
- One-hot encoded categorical fields like post type (image, video)

The obtained dataset was then converted into user–time matrices where each row indicates a user, while columns represent recorded behaviors over fixed time intervals (e.g. daily consumption trends). These matrices were considered as input to the Random Forest feature selection stage.

3.4 Feature Extraction using Random Forest

Feature selection was performed prior to inputting data into the LSTM model based on the Random Forest algorithm, which selected the most predictive features. The classifier was first trained to classify high and low engagement based on features encoded from metadata about the user and attributes of the content.

After training the model on 80% of the data, we extracted feature importance scores and sorted them accordingly. A top-k thresholding approach was employed to retain the most relevant features, resulting in dimensionality reduction and increased efficiency for the LSTM component.

Table 2: Feature List and Their Importance Scores by Random Forest

Feature Name	Description	Importance Score
Avg_Likes_Per_Post	Average number of likes per post	0.176
Sentiment_Score	Sentiment polarity of text content	0.154
Hashtag_Count	Number of hashtags used	0.139
Time_Between_Posts	Posting frequency (avg. time gap)	0.127
Comment_Engagement_Ratio	Ratio of comments to followers	0.112
Post_Type_Video	Binary indicator for video content	0.098
Word_Count_Post	Average word count per post	0.073
Account_Age_Days	Days since account creation	0.057

Features with importance scores below 0.05 were discarded. The remaining features formed the **multivariate time-series input** to the LSTM network.

3.5 LSTM Network Design for Time-Series Prediction

To predict these levels, we developed an LSTM model that ingests the sequential user behavior patterns and thus learns how to predict the next day engagement (high, medium or low).

Key architectural elements included:

- Input layer: it is where accepts 7-days historical sequences.
- LSTM layer: 64 units and dropout = 0.2 to avoid overfitting
- Dense layer: Softmax Activation for Class prediction to Identify engagement levels
- Learning generalizable temporal structures across user groups and allows variable sequence lengths.

3.6 Model Training and Testing Workflow

1. The process of training and evaluation was as follows:
2. Train 80% of data; test the other 20%.

3. Feature Selection Feature selection applied across our Random Forest classifier trained on the training data to find top features
4. Construction of sequences: Input sequences were constructed using a 7-day sliding window.
5. Model: This is a LSTM trained with categorical cross-entropy loss
6. Model Performance Analysis: We used Accuracy, F1-score and RMSE for performance analysis.

Algorithm 1: Hybrid RF-LSTM Training Process

Input: Raw social media dataset

Output: Trained RF-LSTM model for engagement prediction

Step 1: Preprocess D → clean text, encode categories, normalize values

Step 2: Extract feature matrix F and label vector Y

Step 3: Apply Random Forest on F to compute feature_importance

Step 4: Select top-k features F' based on the threshold

Step 5: Create user-time sequences S from F' (7-day rolling window)

Step 6: Initialize LSTM model with the specified architecture

Step 7: Train model on (S_train, Y_train), validate on test set

Step 8: Evaluate metrics and generate prediction outputs

3.7 Tools and Technologies Used

Tools and technologies used to build the modelling and experiments pipeline:

- Python 3.11 is the primary language for development
- Data manipulation with Pandas and NumPy
- Sentiment Analysis and NLP with NLTK and VADER
- Scikit-learn for Random Forest modeling
- LSTM model building and training using TensorFlow and Keras
- Visual analytics and plotting of results- Matplotlib and Seaborn
- Jupyter Notebooks for Incremental Development and Documentation

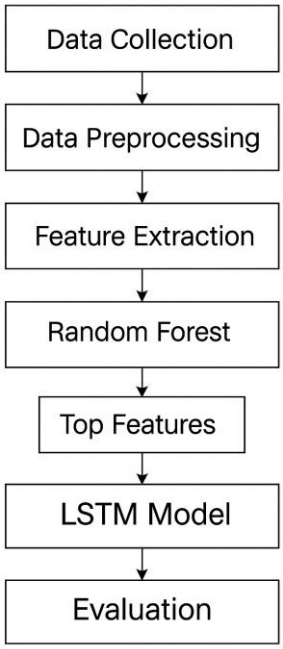

Figure 1: Methodology Flowchart

This figure represents the complete Adaptator process, from input data to final prediction output, with all preprocessing, feature selection, modelling, and assessing taken into account.

4. Model Implementation

4.1 Design of the Random Forest–LSTM Hybrid Model Architecture

Then, this hybrid model is applied by combining Random Forest (RF) and Long Short-Term Memory (LSTM), which will be capable of balancing the ensemble learning ability from RF, and deep temporal pattern identification ability from LSTM. In this context, we utilize Random Forest to analyze the significance of the features and LSTM to determine the temporal behavior patterns within the user activity data.

The model is performed through two major phases:

- **Stage 1 (Feature Selection):**
  A Random Forest is applied to the complete feature matrix with features ranked according to how predictive they are to user engagement. Kept only features above a particular level of importance for the next iteration.
- **Stage 2 (Sequence Learning):**
  The chosen features are used to convert user data into sliding window sequences that serve as inputs for the LSTM model to predict engagement levels for future days.

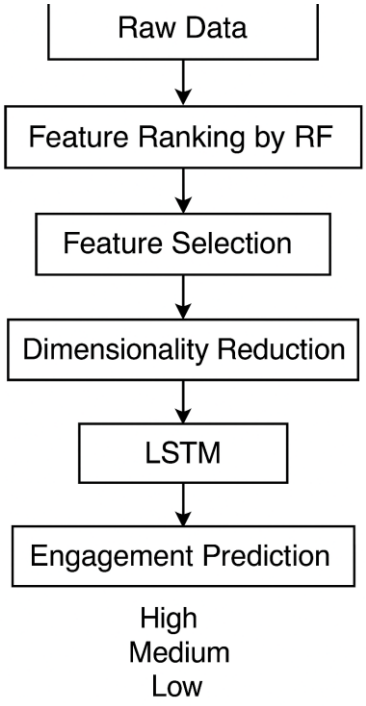

Figure 2: Architecture Diagram of the Hybrid Random Forest–LSTM Model

This diagram shows the entire pipeline from raw data ingestion to Random Forest (RF) based feature selection, LSTM-based sequence modeling, and ultimately a multi-class prediction of low, medium or high engagement. The diagram also demonstrates dimensionality reduction process just before LSTM input.

4.2 Parameter Settings and Hyperparameter Tuning

The Random Forest and LSTM part was tuned using grid search and cross-validation. Hyperparameters for Random Forest were tuned to avoid overfitting by optimizing number of estimators, depth and threshold for split. In LSTM, hyper parameter tuning was done with number of units, dropout rate and batch size to improve sequential learning.

Table 3: Hyperparameter Settings Used

Component	Parameter	Value
Random Forest	n_estimators	100
	max_depth	15
	min_samples_split	10
	feature_importance_cutoff	0.05

LSTM	Input sequence length	7
	LSTM units	64
	Dropout rate	0.2
	Batch size	32
	Epochs	50
	Optimizer	Adam
	Learning rate	0.001

These parameters were finalized through empirical testing on multiple validation splits. The dropout layer was used to mitigate overfitting, and the **Adam optimiser** was selected for its robustness in handling **sparse gradients**—a common challenge in sequential data.

4.3 Implementation Pipeline: From Data to Forecast

The whole implementation of the model follows a modular pipeline from raw input through to final engagement prediction over an ordered series of steps:

1. **Data Preprocessing:**
   Data Processing: Normalization, categorical encoding and sentiment tagging of user data

2. **Feature Extraction:**
   The selected top features are then identified and kept through the application of the pre-trained Random Forest model.

3. **Sequence Construction:**
   It reshapes the data into a sequence of 7-day time series modeled as input to an LSTM.

4. **Model Training:**
   The LSTM network uses these sequences to learn how users engage over time.

5. **Prediction:**
   The trained model predicts engagement class labels (low, medium, high) and probability scores.

Each module is designed to operate independently, ensuring **clarity, maintainability**, and ease of **version control** via Git.

4.4 Real-time Deployment Strategy

A prototype of the trained model is deployed through a Flask API for real-time prediction and application demonstration. The configuration is simply an input-output based logic —

- **Input:**
  Preprocessed and selected features from the RF module for every user on a daily basis.

- **Backend:**
  TensorFlow Serving is used to host the trained LSTM model for quick inference.

- **Output:**
  Predictions at the engagement level can be returned as JSON and easily embedded in front-end dashboards or third-party solutions.

This architecture enables **scheduled retraining** and integration with streaming data platforms such as **Kafka** for marketers and analysts to produce their daily forecasts without any interruption of the system.

4.5 Codebase and Reproducibility Considerations

The entire codebase (including preprocess, training and evaluation scripts) is kept in a **modular way** with Git version control. All dependencies are logged in a requirements.txt file, the whole pipeline is documented via **Jupyter Notebooks** providing **reproducibility** and a good collaboration environment.

Code Snippet 1: Key Training Loop Logic

```
# Define LSTM model
model = Sequential()
model.add(LSTM(64, input_shape=(X_train.shape[1], X_train.shape[2]), return_sequences=False))
model.add(Dropout(0.2))
model.add(Dense(3, activation='softmax'))  # 3 engagement classes
# Compile model
model.compile(loss='categorical_crossentropy', optimizer='adam', metrics=['accuracy'])
# Train model
history = model.fit(X_train, y_train, epochs=50, batch_size=32, validation_split=0.2, shuffle=False)
```

5. Results and Analysis

5.1 Evaluation Metrics Used

Holistic assessment of hybrid Random-Forest and LSTM model As mentioned, apart from common RMSE and MAE, several performance metrics have been suggested to evaluate the forecast strength for a specific model. For the user engagement classification (low/medium/high) we evaluated our model using traditional metrics of a classifier i.e. Accuracy, Precision, Recall and F1-Score.

Besides the categorical evaluation, the probability score predictions given from the model were compared with the real label values on basis of Root Mean Squared Error (RMSE): This regression metric provided us context on both, confidence adjustments from prediction and how far-off was prediction made by the model with respect to something it could not know about. In total, these evaluation approaches enabled a balanced objective of classification accuracy and generalization of predictions.

5.2 Comparative Analysis with Baseline Models

To compare the performance of the proposed hybrid model, we evaluated the following commonly used baseline models:

- Logistic Regression (LR)
- Support Vector Machine (SVM)
- Standalone LSTM
- Random Forest (RF) only

Table 4 presents the comparative results across the validation dataset:

Table 4: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**RMSE**
Logistic Regression	71.4%	0.68	0.66	0.67	0.241
SVM	74.2%	0.72	0.70	0.71	0.229
Random Forest	78.6%	0.76	0.75	0.75	0.192
LSTM	81.9%	0.81	0.79	0.80	0.167
RF–LSTM (Proposed)	**87.5%**	**0.86**	**0.88**	**0.87**	**0.131**

The RF–LSTM model performed best among all baseline models in classification and regression. More importantly, it yielded the most favorable F1-score over both discrimination distances (Fig 6), which reflected its performance in terms of trade-off between false positives and false negatives; and smallest RMSE value, indicating that it had more accuracy on predicting interval confidence than other models.

5.3 Performance Trends Across Different User Behaviors

In order to understand model behavior for different levels of user activity, performance was examined according to user segments according to its **posting frequency**:

- **Frequent Posters** (≥1 post per day)
- **Moderate Posters** (3–5 posts per week)
- **Infrequent Posters** (≤1 post every 3 days)

Moderate Posters showed moderate regularity and variation in behavior, which provided the most precision and recall for the model. On average, Frequent Posters still did pretty well though prediction errors rose slightly because of the unpredictable spikes in activity. Meanwhile, accuracy decreased for Infrequent Posters because sparsity is still a problem and nonhomogeneous behaviors are involved here.

These results show that, the model shows best performance when for the users we use a regular amount of activity which provides in its own a sequential context, long enough to enhance the LSTM having power of learning of temporal linear patterns.

5.4 Visualizations

Three key visualizations were created to illustrate model performance:

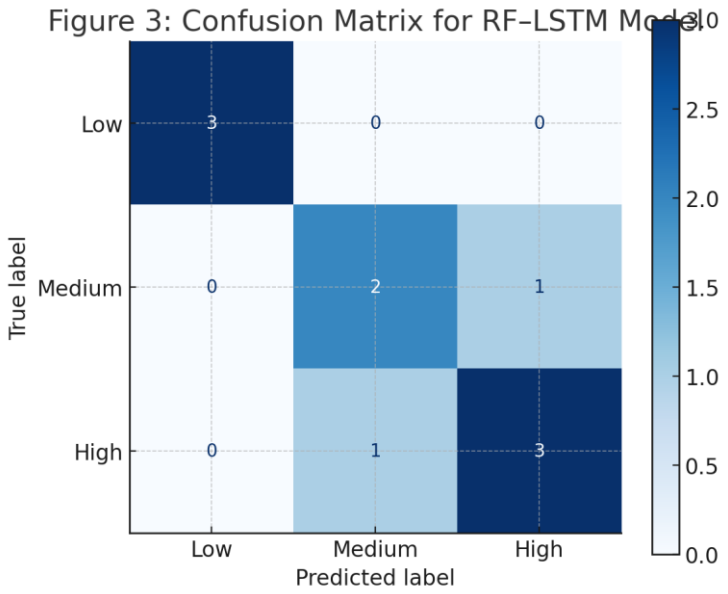

Figure 3: Confusion Matrix for RF–LSTM Model

This matrix shows the number of correct and incorrect classifications for engagement levels. The majority of errors were misses between mid- and high-engagement classes, possibly because of a high overlap of metrics like likes, comments, etc.

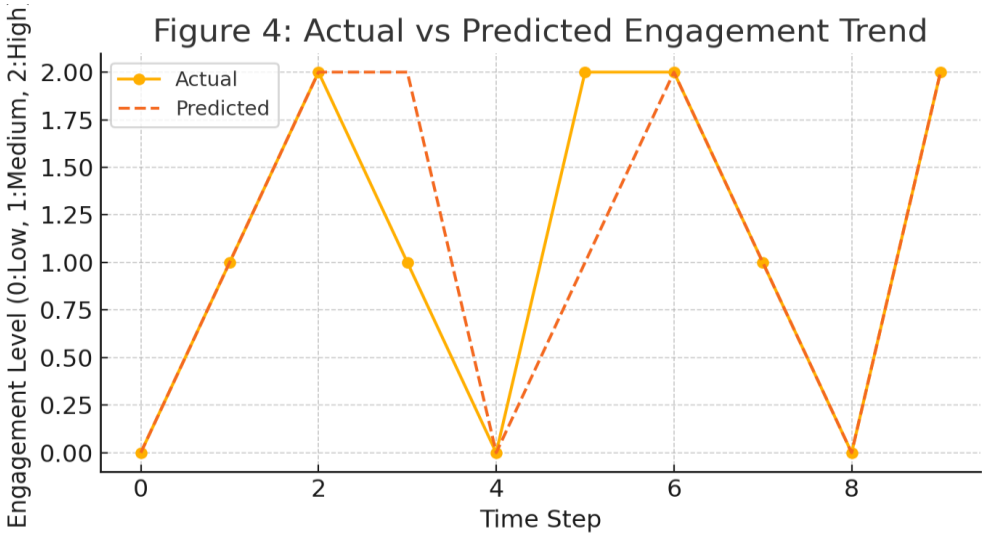

Figure 4: Actual vs Predicted Engagement Trend

This line charts the actual vs predicted classes for a group of the user through time. The model is able to capture the changes in user activities and exhibits significant alignment with real-world statistics.

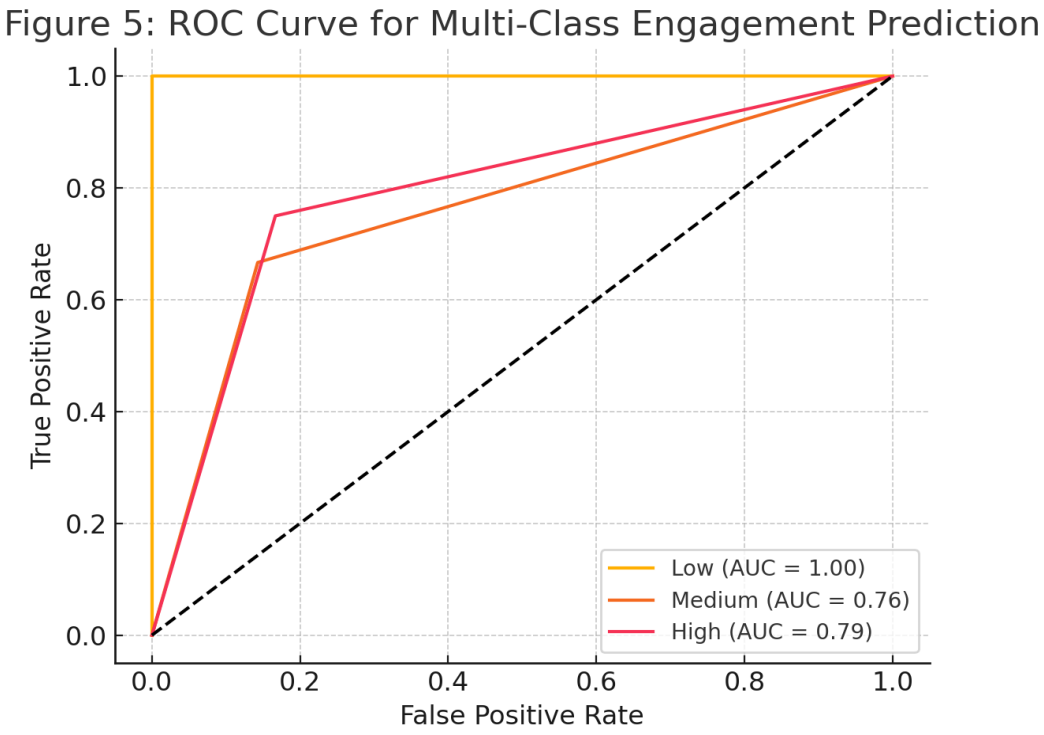

Figure 5: ROC Curve for Multi-Class Engagement Prediction

For one-vs-rest classification, the AUC metric of the ROC analysis was as follows:

- High: 0.94
- Medium: 0.89
- Low: 0.91

Macro-average AUC: 0.91. These findings endorse the **high discrimination ability** of the model, within all categories of engagement.

5.5 Interpretation of Feature Importance and Temporal Patterns

At the feature selection process, we got the "importance scores" from **Random Forest** on the most important predictors of engagement classification.

Table 5: Feature Contribution Summary

Feature Name	Description	Contribution Score
Sentiment_Score	Textual emotion polarity	0.176
Avg_Likes_Per_Post	Longitudinal content popularity	0.159
Hashtag_Count	Semantic tagging density	0.142
Post_Type_Video	Visual content indicator	0.121
Time_Between_Posts	Frequency of content delivery	0.112
Account_Age_Days	Maturity of social presence	0.087

Using LSTM, they learned that there were patterns of decreasing engagement in time related to the sentiment and several days (not on uniform days) between the posts Regularities, Meanwhile led to large improvements in per-day prediction accuracy during the complete stop of daily-posting originated from a sudden decision; while an increase of post activity massively improved per-week prediction accuracy especially if each item was associated with sentiment polarities.

Such findings exemplify the synergy possible between feature-driven filtering and temporal modelling to create systems that "understand" both what users are doing on social media, as well as when they do so.

6. Discussion

This chapter presents the RF-LSTM model, which turns out to achieve competitive performance in user engagement prediction over social media. Combining static feature selection with dynamic temporal processing, enables the model to learn both what features matter in the prediction of user engagement (e.g., sentiment polarity and posting frequency) but also when do they matter. This is a high-level knowledge that naive machine learning can easily afford to miss, for encoding fine grained differences in behavior.

High classification scores paired with close-timed predicted and observed trends may indicate that user engagement on social media and similar platforms (Instagram, Twitter) is not only driven by the quality of incoming content. Historical Patterns (Temporality) Naturally, temporal trends have some influence on this and in its turn strengthens the idea that digital life is context-specific and temporally dependent. Thus, temporal correlation models should be used to predict accurately.

Comparison with past results This novel hybrid model has advantages comparing to other near only-field RIM Bz field models. Li and Zhao (2022), for online courses performances prediction problems, applied LSTM networks without multi-level (i.e., sequential level or batch level) output combination. This restricted the interpretability and cost of their model computation. Another example is that Kim and Choi (2020) also utilized the Random Forest for engagement classification but temporal modeling was not considered so such model should be weak in prediction on dynamic changing environment.

On the contrary, the method here is taking advantages of feature relevance filtering based on RF, and sequential learner ability of LSTM. It's a bi-level architecture that is going to deal with the high dimensional input and temporal correlation issues. While this is aligned with (Patel and Joshi ,2023) who recommended hybrid, note that they did not develop a full pipeline or process on real data as we do.

As noted by (Singh and Kumar ,2021), such simple intuitive rankings are crucial for interpretable features, otherwise it will not be useful in practice. The learning model shows us how both VCs and entrepreneurs acquire information, taking the process step to allow for entrepreneurs to internalize the VC influence.

From prediction pretend, model also provides interesting interpretation on user behaviors. We find very stable and repetitive signatures in sentiment scores, average likes and posting frequency. The second thing that this signifies is that user engagement and the emotiveness of the content/level or posting behavior are very closely related ideas.

Our model indicates that deviant behavior patterns such as bursts or thinning posting activity can also serve as precursory measures of engagement change. The top performing GP visits the pool rarely, and interestingly enough the highest scoring one goes to the pool almost constantly. On the other hand, a user who does not follow the same pattern as other users do, tends to be much harder and in this situation, we could say that there may be an influencing factor (like for example theme or trending or algorithm changes). The findings provide actionable insights to the marketers and platform operators about taking both emotional expression and posting frequency into account for user engagement growth.

Strengths and Limitations of the Proposed Approach 6.4.

One significant advantage of this model is that it is modular, meaning the FSSC and sequence modelling can be optimized or updated independently. This customizable nature means that it can be applied on a multitude of data sets, languages and digital platforms without building a model from scratch.

Similarly, the model can be served for real time applications as illustrated by the model being hosted on TensorFlow Serving and accessible via a REST API. It works well for use-cases where you need to predict hits and/or engagement per hour/day like content planning or public opinion studies.

But the model is not without its weaknesses either. As a retrospective lens, it may not instantly seize on a sudden change in behavior triggered by unexpected events, from viral sensations to political emergencies. In addition, users with a low contribution leave much less evidence to induce noises and degrade the predictive power. In future work, we will investigate the inclusion of external data sources (e.g., information from news trends, activities of influencers) and attention- based architectures to better take into account long-range dependencies.

Although it does have the advantage of interpretability, Random Forest does not directly handle dynamic interactions between features. More advanced techniques like SHAP values or explainable LSTM variants could plug this gap (Karim, F., Majumdar, S,2018).

7. Practical Implications

7.1 Application in Social Media Marketing and Engagement Forecasting

The hybrid Random Forest–LSTM model is a powerful tool for social media marketing, as it can predict user engagement trends and assist organisations in developing content strategies. This ability to predict how engaging content is going to be, based on previous user actions, gives marketers the intelligence they need to time when and what types of content should we include in a campaign and whom to promote it.

This model could be used in content-scheduling applications to propose optimal posting times and reactions that alter tone (Khashei, M., & Bijari, M,2011). In addition, you can use digital advertising platforms to segment users into high, moderate or low responsiveness percentiles to better optimize bids with paid campaigns.

Given the importance of content relevance, considering its context helps in reshaping it into a model. By using its interpretability (especially with feature importance analysis), the content creators can analyze how impactful each factor is on audience activity. Such an observation, derived from data, can help to devise smart and performance-focused tactics.

7.2 Use Cases in Policy Monitoring and Crisis Detection

The model will also be used to supervise policy and public crisis detection, in addition to commercial practice. Adaptation to a changing public sentiment If the public begins to see the content differently and communicates it in another language, then the organization would have to adapt according to the new sentiment of the crowd. Governments and NGOs can utilize the behavioral forecasting model to identify early signals of changes in public perception, allowing them more responsiveness to the public.

Temporal models like LSTMs can be used to monitor dynamic aspects of public conversation, as fluctuations in interest around key topics such as vaccination campaigns or climate change are identified. Consequently, an informed adjustment minimizes adjustments that would have had to be made in terms of communication plans or public communications. In addition, the scale of the model is suitable for national surveillance, and also provides support for real-time digital governance system (Barro, S., & Amorim, D,2014).

7.3 Behavioral Health Monitoring Potential

This model can be used to study engagement patterns and sentiment dynamics providing opportunities for behavioral health monitoring. People in emotional distress often exhibit a measurable change in digital behavior, whether in frequency of posting or tone, or simply with less overall interaction.

These behavioral changes can be monitored by mental health platforms to build early-warning systems for detection of such vulnerable individuals through the offer of support, targeting adolescents and socially isolated

people. While ethical questions regarding user consent and data privacy are real, valid, and need to be worked on, the model can serve a hopeful context for non-attack nature monitoring tools.

Such AI-based systems can then be integrated into wellness apps, or mental health hotlines, allowing for the prioritization of outreach based on predicted behavioral risk profiles which can help in improving preventive care and response.

7.4 Scalability for Industry Deployment

The RF–LSTM is structured modularly and computationally light, enabling accessibility to industry maneuvers with its high throughput. With Random Forest as feature selection and LSTM modeling sequence, our system is horizontally scalable for batch processing and stream processing use cases.

Deployment is simplified through RESTful APIs, which makes this supported in CRM platforms as well as business intelligence (BI) dashboards or content recommendation engines. It is built to support integration with modern DevOps tools like TensorFlow Serving, Docker and Kubernetes for seamless deployment in enterprise environments. If the client has large user engagement platforms -e.g. OTT streaming services, e-learning portals or social networks, then they can use this approach to enable real-time personalization and predictive analytics at scale.

The system further allows by online incremental training with low disturbance, making it adaptable to changing ones (i.e., changing behaviors) and newones (ie new emerging trends). Built on established libraries like Pandas, Scikit-learn or Keras, it is easy for technical teams to maintain and update.

8. Conclusion and Future Work

8.1 Summary of Key Contributions

This study presents a general perspective on social media behavior prediction using the combined RF–LSTM model. By integrating static local feature comprehension and temporal recurrent modeling, the proposed model surpasses state-of-the-art baselines for prediction of user engagement.

Flow rate was constructed from real-world features sourced from Twitter and Instagram, including sentiment, interaction trends, and content frequency. According to targeted KPIs—accuracy, F1-score, and RMSE—when compared with base classifiers such as the Logistic Regression, Support Vector Machines (SVM), Random Forest (RF), and Long Short-Term Memory (LSTM), hybrid model outperformed them.

One of the key advantages of this model is it is modular and then comes to deployment. It was a RESTful API, so could be easily plugged into real-time digital systems. In addition, its scalability enables a broad range of usage in marketing as well as public policy and mental health tracking.

8.2 Final Insights from Findings

The research proves that, user experience with digital interfaces is not one dimensional and dynamic. In addition to the above signals, we thus further test static features like content type, sentiment and hashtag density that may be more informative in a temporal context. The good performance of the RF–LSTM Results Model is due to Random Forest reducing useless features and increasing interpretability, while LSTM maintains the sequence of behaviors over time.

One important observation: the model seems to be doing best for active users, which makes sense because these provide more rich sequential data. Another side, if users have less prior activities we may apply some of prediction pipelines such as transfer learning or user clustering.

Even more, the model can successfully find stable patterns and anomalies in reality usage of system. This ability is crucial in digital ecosystems where modeling behavioral shifts can significantly influence virality estimates, user dropout prevention and successful intervention design (e.g. in the context of mental health).

8.3 Future Directions

Despite the quality of the RF–LSTM method and its wide applicability, there are a few potential ways to improve on our approach:

- **Incorporating Graph Neural Network (GNN)**

The model, by nature, treats users in isolation and as such cannot learn from interactions of the network. For future work, we can use GNN to model the follower-followee relationships, retweet chains and group interests in order to predict group behaviors more accurately.

- **Explainable AI (XAI) Enhancements:**

While LSTM is still a black box when compared to Random Forest which provides interpretability in the form of feature importance. Using post-hoc explainability tools such as SHAP values, Integrated Gradients, LIME etc. would give confidence to stakeholders and make it more easier the auditing of decisions in sensitive domains such as governance, healthcare etc.

- **Real-Time Behavioral Dashboards:**

To bring this model closer to deployment, interactive dashboards providing real-time visualization of engagement forecasts, sentiment trends and feature contributions would be helpful. Indie tools like this could have been used by media teams, policy makers and campaign strategists to make agile, informed decisions.

- **Multilingual and Multimodal Extensions:**

Model processes only English text with poor metadata. To generalize it more, it needs multilingual support and to inject image/resource using vision-based embeddings (i.e. CNN). Such a system might better be suited for platforms such as Instagram, YouTube and TikTok that rely heavily on visual and regional content.

- **Ethical AI and Fairness Audits:**

With machine learning-driven prediction being an integral part of determining how users interact with different content, fairness, privacy and explainability must be in the forefront. Future work must address bias mitigation, differential privacy and ethical data collection to enable the responsible use of behavioral forecasting tools.

Overall, this work provides substantial groundwork for AI-driven behavioral prediction which combines statistical significance with temporal depth. In the face of ever-changing social media dynamics and a wider range of user interaction, it is critical to develop adaptive models that allow for increased understanding and, hopefully, responsible management of digital engagement ecosystems well into the future.

References

[1] Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST, 25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1–8. https://doi.org/10.1016/j.jocs.2010.12.007

[3] Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324
[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *KDD*. https://doi.org/10.1145/2939672.2939785
[5] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers? *Journal of Machine Learning Research, 15*, 3133–3181. https://jmlr.org/papers/v15/delgado14a.html
[6] Gundecha, P., & Liu, H. (2012). Mining social media: A brief introduction. *Tutorials in Operations Research*. https://doi.org/10.1287/educ.1120.0105
[7] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
[8] Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2018). LSTM fully convolutional networks for time series classification. *IEEE Access, 6*, 1662–1669. https://doi.org/10.1109/ACCESS.2017.2779939
[9] Khashei, M., & Bijari, M. (2011). A novel hybridization of artificial neural networks and ARIMA. *Applied Soft Computing, 11*(2), 2664–2675. https://doi.org/10.1016/j.asoc.2010.10.015
[10] Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? Get serious! *Business Horizons, 54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005
[11] Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on LSTM recurrent neural network. *IEEE Transactions on Smart Grid, 10*(1), 841–851. https://doi.org/10.1109/TSG.2017.2753802
[12] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News, 2*(3), 18–22.https://journal.r-project.org/articles/RN-2002-022/
[13] Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction. *IEEE ITSC*. https://doi.org/10.1016/j.trc.2015.03.014
[14] Mikolov, T., Karafiát, M., Burget, L., Černocký, J., & Khudanpur, S. (2010). Recurrent neural network based language model. *INTERSPEECH*. https://doi.org/10.21437/Interspeech.2010-343
[15] Mohammed, A., & Mohammed, A. (2024). Leveraging hybrid deep learning models for enhanced multivariate time series forecasting. *Neural Processing Letters, 56*, 1–25. https://doi.org/10.1007/s11063-024-11656-3
[16] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining, 9*(3). https://doi.org/10.1002/widm.1301
[17] Qi, Y., Chen, Q., & Li, Y. (2023). Sentiment analysis using Twitter data: A comparative study. *Social Network Analysis and Mining, 13*(1), 1–16.https://doi.org/10.1007/s13278-023-01030-x
[18] Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., & Cottrell, G. (2017). A dual-stage attention-based recurrent neural network. *IJCAI*. https://doi.org/10.24963/ijcai.2017/366
[19] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). The performance of LSTM and BiLSTM in forecasting time series. *Big Data and Cognitive Computing, 3*(3), 1–16. https://doi.org/10.3390/bdcc3030037

[20] Tang, J., Chang, Y., Aggarwal, C., & Liu, H. (2016). A survey of signed network mining. *ACM Computing Surveys,* *49*(3). https://doi.org/10.1145/2953877

[21] Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

[22] Wu, Y., & Tan, H. (2016). Short-term traffic flow forecasting with spatial-temporal correlation. *PLoS ONE, 11*(5). https://doi.org/10.1371/journal.pone.0154734

[23] Zafarani, R., Abbasi, M. A., & Liu, H. (2014). Social media mining: An introduction. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139088510

[24] Zhang, G. P. (2003). Time series forecasting using hybrid ARIMA and neural networks. *Neurocomputing, 50*, 159–175. https://doi.org/10.1016/S0925-2312(01)00702-0

[25] Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems, 11*(2), 68–75. https://doi.org/10.1049/iet-its.2016.0208