JSUCIT

Journal of Shaqra University for Computing and Information Technology
ISSN 2961-4279 (Print) || ISSN 2961-4287 (Online)

AI-Based Modeling of Social Media Engagement Using Sentiment Analysis, Clustering, and Time-Series Forecasting

Mohamed Alshalaan[1]*
Department of Computer Science, College of Computing and Information Technology,
Shaqra University Riyadh Province, Saudi Arabia
Corresponding Author Email: m.alshalaan@su.edu.sa

Abstract Social platforms generate vast, fast, and noisy streams of user discourse. Institutions and brands alike increasingly need actionable intelligence about when, why, and how people engage with content. This paper presents a practical end-to-end framework that combines multilingual sentiment analysis (Transformers), unsupervised engagement clustering (K-Means), and temporal forecasting (LSTM) to detect and interpret engagement patterns. We operationalize and evaluate the framework on a synthetically generated dataset of tweets collected around the education conversation in the Kingdom of Saudi Arabia (#EduKSA). The pipeline yields distinct engagement regimes (low/medium/high), shows strong correlation between positive sentiment and higher interaction, and produces short-horizon forecasts useful for timing outreach. We provide inline figures and tables and discuss implications for institutional communication and knowledge sharing.

***Index Terms*— Social media analytics, engagement prediction, sentiment analysis, K-means clustering, DBSCAN, Long Short-Term Memory (LSTM), time-series forecasting, explainable AI, SHAP, natural language processing.**

I. Introduction

Several social media platforms have been working on the transformation of communication and information sharing for a long time. The platforms can generate a huge amount of data that contains comments, sharing, tweeting, and various other psychological and economical dynamics. It is a very critical phenomenon to understand the pattern of the engagement of the users, especially when it comes to business marketing and brand reputations. In general, monitoring the opinion given by various policymakers and information that is spreading across the globe must be studied in a very dynamic way. There are various traditional methods that are accounting for this activity, and it is expected to scale in complexity very fast. The use of artificial intelligence, including several machine learning algorithms empowered by natural language processing, is one of the revolutionary techniques that is used for understanding the dynamics of social media analytics [1]. Without the use of manual coding of systems, artificial intelligence can scale the unstructured data and identify various patterns with the help of forecasting the behavior [2]. In this case the engagement of the users is complex, and a very high probability is accepted towards politeness or comment scrolling given by the users. Advanced approaches are required for the integration of various modeling techniques, analysis of the language, as well as prediction of the behavior to identify the sentiments [3, 4].

In this study we try to propose an AI-enabled framework that can provide comprehensive tasks such as detection and interpretation of the engagement patterns of several users that are integrated into various modules. The collection of data along with the processing and analysis of sentiments, including the clustering and predictive modeling, is helpful for the identification and forecasting of the information using deep learning techniques. We made use of a synthetically generated dataset like the tweets produced on the X platform as a case study for the demonstration on modular architecture. This comprises of an analysis of the specific engagement pattern to provide the contribution towards the understanding of various platforms and data sets. The work focused mainly upon descriptive matrices that are provided for sentiment analysis, as our approach tends to combine various elements in the form of a unified platform pipeline. There is a strong significance in practical sectors that include business, where the AI-driven analysis will be helpful towards the resonance of their target audience that are able to depend upon market investments [5–9]. In the field of politics as well as the creation of public policies, the opinion of people is required, and misinformation may lead to issues concerned with the political and social mobilization movements. There are various campaigns that involve the use of several activities such as vaccination, engagement of the general population, economic reforms, public assurance, well-known topics and articles, etc. Nonprofit organizations and educational organizations are one of the key factors that provides maximum reach and impact on the population, which is under study for any specific task [10–14]. In these applications the engagement of the users enables the activity and adjustment towards the detection of various concerning trends that are coordinated across the world. In a framework that we have identified contains working across these types of analyses that require the preprocessing of the data, feature engineering, classification of the information obtained, and forecasting various stages depending upon the valuable and refinable reproduction of new methodologies. All the modules that are concerned in this framework tend towards the identification and creation of various loopholes.

Research Gap and Research Questions: There has been several prior works conducted in the same field who have separately addressed sentiment classification, engagement clustering, and virality forecasting. But in general, most of them integrate these three components into a single, reproducible pipeline relative to simple baseline evaluation, reporting of hyperparameters and computational costs. This produces a gap that to limit practitioners and reviewers to judge in case if the added complexity of deep learning is justified for a given engagement-prediction task or not. Motivated by the gaps we tend to identify the response for the following gaps:

- Can a modular pipeline connect the sentiment analysis, unsupervised clustering, and LSTM-based forecasting for social media?
- Does an LSTM forecaster provide a measurable accuracy improvement over simpler baselines?
- Which features (sentiment, timing, content, or audience metadata) contribute most to predicted engagement?

The basic framework proposed in this study engages the decomposition of the analysis with the help of preprocessing, feature engineering, and classification to involve forecasting of the information based on the sentiments and emotions of the users. In this architecture the integration of several features is done with the help of TF-IDF (Term Frequency–Inverse Document Frequency). For the interoperability and interpretation of various semantics that are concerned with the transformer-based embedding of linguistic sarcasm context-independent meaning of the information processed. The second layer combines the use of unsupervised learning with the help of supervised classification to identify various engagement regimes and find out the prediction referring to the engagement levels of the users. In the final layer, the employability of LSTM and various transformer-based architectures is done to identify and capture the information for various sequences and semantic richness. However, in the final stage of processing, the prioritization for explainability is done with the help of SHAP (SHapley Additive exPlanations) values that are equally responsible for the identification, analysis, practitioner enablement, and understanding of the decisions taken with respect to this information.

The social media engagement of the users involves the integration of personal data along with the decision for ethical consideration, having privacy, algorithmic bias, and manipulative potential with various rigorous safeguards and safety standards. The methodology that is proposed in this scheme concerns the prioritization of public data under the PII removal and GDPR (General Data Protection Regulation) compliance for various individual-level analyses. In this article, the intersection of artificial intelligence along with social media analysis tends towards the identification of behavioral results. The integrated framework that is proposed in this study combines sentiment analysis and clustering with the help of deep learning and interpretation of ethical rigor to identify the gaps in various existing methods. The article also contributes to several computational social science actionable items that are used for data-driven decision-making. However, in our findings, we illustrate that AI methodologies can be used for proper designing and evaluation of digital engagement with the audience. In this article section two includes reviews of the engagement, their conceptualization, and the AI methodology used. Section 3 enables the methodology that is presented by us in this article. Section 4 tries to present several systems and architectures that are used for the mathematical foundation used in this article. Finally, Sections 5, 6, and 7 provide detailed experimental analysis and results along with the discussion. Finally, section eight concludes the implications and the future directions that are expected in this study.

II. LITERATURE REVIEW

Social media engagement includes user interactions such as likes, comments, shares, retweets, and passive behaviors. These engagements can reflect psychological, social, and economic dynamics of the user mind and a complete community under consideration. However, the conceptualization of engagement spans towards the attention that is devoted towards the content, finding the effective response on an emotional basis, and identification of visible actions towards the behavioral dimensions of the users. The engagement is usually inherited with the help of support, politeness, or an acknowledgement given by a user based on his activity. The sharing of information indicates the endorsement of any article or information towards a proper critique that is presented towards a piece of information.

The ambiguities that arrive in various cultural contexts may necessarily be identified with the help of analytical methods, including various linguistic analyses along with natural network-based modeling for behavioral prediction systems. Several social media researchers in the previous studies depend upon the identification of descriptive statistics in terms of manual coding for various datasets, which were smaller in size. When the platforms are scaled to several billions of activities performed per day and the computational methods are improved, the advent of big data analytics started in the early 2010s with the automation of the classification along with the help of machine learning algorithms and several well-known algorithms like Support Vector Machines (SVMs), Naïve Bayes, etc. [15] These algorithms were able to identify various reasonable outcomes dependent upon the simple task; however, they struggled with various multilingual issues, and sentiment analysis was lacking. The use of deep learning revolutionizes the analysis of the social media platform and its related content. The convolutional neural networks, along with the help of long short-term memory networks, were able to identify and capture the sequence of various textual information that was used for the traditional methods.

It started to provide much better results as compared to the previous heuristics used in the disregards. Several transformer-based architectures suggested by GPT and BERT were able to make use of advanced natural language processing techniques to identify the modeling context along with the linguistic nuance to overcome the detection of sarcasm along with hate speech and virality prediction techniques [16]. These models were able to identify the foundation of various engagement analytics patterns that were taken care of by several systems suggested by various authors. The indication of engagement depends on multiple factors and how they are interpreted. The likes and reactions of various proxy information are used for emotional responses and can be able to comment on various qualitative data, which is used for sentiment as well as logic modeling. The sharing and the retweeting of the information by various users possess the content of diffusion patterns across the entire network. Several researchers have been increasingly able to employ various measures that are used for the combination of multiple pieces of information that are accepted from various platforms. The interaction refers to the identification of the engagement for various clusters that are done by the topologies, like liking, commenting, and amplification of the information, which is visible differently in terms of the influencing content.

The predictive model is used for engagement analysis, and it is helpful based on a regression approach to identify the engagement for the nonlinear relationship of information that is posted at different times, in different words, and with various hashtags. Several machine learning algorithms, such as random forest, neural networks, or gradient boost, will be used for feature interactions in several research works given by authors. However, the use of deep learning approaches will be very helpful to incorporate the text content along with sentiment analysis and identification of the network topology for engagement trajectories. The temporal analysis of the information that is done by early engagements with the help of long-term viral information adds to identifying various predictive models to improve the accuracy more significantly for finding out a pattern in the information that is available [17]. The AI-driven engagement analysis can comply with several domains; for example, in marketing, the emotional analysis is more relevant as compared to the content engagement. Domain-specific personalization systems on social platforms, including travel planning recommenders, further demonstrate this trend [18].

The political analysis for any specific platform reveals the various digital campaigns that are done and the coordinated behavior of the people across the community towards this opinion. Various public health levers were also identified towards the assessment of effective and vaccine statements, and recent work also applies social-media AI signals to mental health risk detection [19]. There are several sociobiological research studies, social capital researchers, activist research studies, etc. that make use of AI to identify various community echo chambers. Ethical considerations are one of the most important aspects of this research, in which privacy concerns are used to analyze data based on the algorithmic bias towards the skewed data set that is used in this study. For various predictive models, sensitive decisions, including politics or health, demand a transparent nature of the algorithm, which is running at the back end. Various researchers usually employ several ethical frameworks that are used for balancing the rights of the user as well as social responsibilities depending upon privacy elevation, including explainability of the AI recommendation behavior on social platforms [20].

Various emerging trends such as multimodal analysis that comprises text, images, audio, or video across the class platform are used to analyze various interconnected engagement ecosystems, including generative AI content ecosystems and learner responses in social media contexts [21, 22]. The real-time engagement of the data with the help of streaming analysis as well as explainable AI is used for clarification of various features that can be helpful to contribute to the predictions. The development of help for better techniques and technology to understand the engagement of the audience is more required as compared to various accountability challenges. In this study we try to advance our engagement pattern and identification of various integrations of sentiment classification dependent on behavioral engineering classification and the forecasting of the data, which is transparent and ethically grounded in the framework.

III. Methodology

This study provides an integrated framework that makes use of artificial intelligence by balancing various computational information for interpretability. The use of a ma- chine learning algorithm integrated with natural language processing for the detection of various engagement patterns related to emotional and sentimental analysis of the users is presented in this study. The core components of this methodology involve data collection, processing of the information, feature engineering, clustering of the data, classification of the various identified data sets, and use of deep learning forecasting and predictive analysis of the information. We are going to present step by step the various components in this methodology.

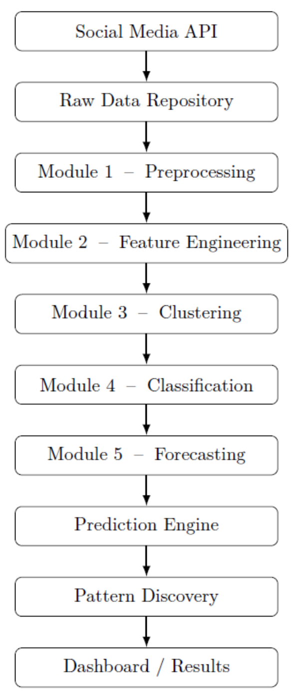

Figure 1. Integrated five-module system architecture

Data Collection: The core idea of the framework suggested is to identify the analytics for sentiments and to identify this, we curated a synthetic dataset comprising of the data posted by users. All results are computed on a single pilot dataset just for the purpose of demonstrating the framework rather than as generalizable population-level findings. In general, the broader validation across platforms, languages, and topics is needed for this work. The sentiment classifier refers to the reduced precision/recall on negative sentiment due to informal language, sarcasm, and context-dependent cues. The dataset is specified for the specific platform and topic for the education discourse in Saudi Arabia as a case study, so the results may not transfer directly to other domains or cultural contexts as far as the information dataset is concerned. The metadata of the authors, along with the conversational structure, was identified and includes follower number, account IDs, and age. A MongoDB-stored unstructured JSON template is used for the storage of data that has a rate limitation, which is managed by various unsynchronized calls.

Pre-Processing: The raw data that is collected from the social media context is screened with the help of NLTK pipelines that are used for the tokenization and careful removal of the words that are preserving various instance-indicative terms (I, we, they, etc.). Preprocessing pipeline is a structured, automated sequence of data-cleaning and transformation steps that converts raw, messy data into a clean format optimized for machine learning models or data analysis. UTF-8 encoded emoji along with the conversion of an equivalent sentiment that was done for this data, which is parsed from the engagement indicators. Several hashtags that were also integrated in the information were integrated in the JSON file. The links and the URL that were mentioned by the user were screened for relevancy and reply to chains. Several user-level features, including the ratio of followers, the frequency of posting social media information, the trends in the temporal area, and network-centralized stability, were also engineered to identify the behavior as well as the depth of the service- level sentiments of the user.

Feature Engineering: The features that were employed in this activity comprise the three-level identification and representation of data. In the beginning the TF-IDF vectors were interpreted, and the next level leads to the use of Word2Vec / GloVe. To maintain semantic relations for various contextual linguistic information. This sentiment is often qualified as VADER for the lexicon-based information and NRC for the emotional lexicon basis. These features include the modeling of various thematic clusters that are responsible for the community detection of the information along with the structural in- fluence. All the features were normalized, and the reduction of the dimensionality was done with the help of PCA / UMAP. This approach for feature engineering is used to get domain knowledge, mathematics, and statistics to select, transform, and create meaningful inputs (called features) from raw data so that machine learning models can learn patterns and make better predictions.

Modeling Approaches: The unsupervised clustering refers to the identification of natural engagement patterns that include positive reinforcement, adversarial deployment, and passive communities. We take the help of K-means and DBSCAN algorithms to achieve this framework since this is a learning algorithm that groups un-labelled data into distinct clusters. The supervised classification for the prediction of the engagement of various users was done with the help of logistic regression since it uses a linear model for predicting probabilities and the random forest algorithm since it ensemble method that combines many decision trees to capture non-linear patterns. We have also included the use of the XGBoost algorithm since it helps to deliver industry-leading accuracy, speed, and scalability on tabular data from the temporal dataset that we have for the analysis in this study. In this deep learning, the LSTMs were used to capture the information and the sequence of information that is fed into the architecture for semantic richness; adaptive and emotion-aware engagement systems provide an additional practical direction for real-time user response modeling [22 - 24]. After the achievement of this architecture, the transfer of the learning information from the model was sent to the engagement prediction pattern.

Evaluation: The implementation was done with the help of Python libraries that comprise several modules and are reproducible for the experimentation with the help of Jupyter notebooks. The GPU required for the achievement of parallelization of the data, which is under consideration with the help of deep learning training, is done by using these Python libraries in the Jupiter notebook.

Ethical Safeguard: The publicly available data was collected to ensure several ethical rights, and personally, the identification of the information is removed, especially the user IDs or hashed data from the publicly available data. This methodology ensures that the regulations for API utilization and the GDPR principle for the minimization of the data for an individual are maintained. This data was analyzed in a way in which the collective and individual level profiling or stigmatization was not done. The design of the information framework that is created comprises an iterative methodology that includes the clustering of the information and classification based on the deep learning refinement. In this way, the cycle that is approached for pragmatic philosophies was taken into con- sideration to ensure that a flexible adoption of the data from the social media dynamics was maintained to ensure methodological rigor and transparency as well as the ethical responsibility of an individual in the framework

IV. MATHEMATICAL MODELLING

In this section we propose a theoretical formalization for future empirical validation of the results from the study. The methodology is iterative in nature. Clustering insights from the analytical model informs the classification refinement and the classification of the results guide deep learning model development. This cyclical approach, grounded in pragmatic philosophy, ensures flexible adaptation to evolving social media dynamics while maintaining rigor, transparency, and ethical responsibility

A. Derivatives

To formalize the analytical framework used to detect engagement patterns across social media platforms, we introduce a mathematical model that characterizes user interactions, temporal behavior, and sentiment–engagement coupling. Let a dataset of posts be represented as

$$D = \{(x_i, t_i, s_i, e_i)\}_{i=1}^{N},$$

where each post x_i is associated with a timestamp t_i, sentiment score s_i, and an engagement value e_i composed of likes, comments, retweets, and impressions. The purpose of this model is to capture (1) the latent structure governing engagement magnitudes, (2) the temporal evolution of these interactions, and (3) the probabilistic relationship between sentiment and user responsiveness.

1) Engagement as a Multivariate Response Variable

We decompose engagement into its constituent observable metrics:

$$e_i = (l_i, c_i, r_i, v_i),$$

where l_i denotes likes, c_i comments, r_i retweets/shares, and v_i views. Since these metrics typically follow heavy-tailed distributions, we apply log-transformation:

$$\tilde{e}_i = \log(1 + e_i),$$

allowing us to model them using a multivariate linear-nonlinear hybrid model. The expected engagement for post i can be written as:

$$\mathbb{E}[\tilde{e}_i] = \beta_0 + \beta_s s_i + \beta_t f(t_i) + \beta_x g(x_i),$$

where $f(t)$ captures circadian and weekly cycles, and $g(x)$ denotes learned textual features extracted from embeddings.

2) Temporal Dynamics Model

Engagement evolves over time as a function of prior interactions. We define a discrete-time temporal process:

$$E(t) = E(t-1) + \Delta(t),$$

where $\Delta(t)$ represents the incremental engagement driven by post visibility and user activity. To model $\Delta(t)$, we use:

$$\Delta(t) = \alpha_1 A(t) + \alpha_2 S(t) + \epsilon_t,$$

where $A(t)$ is the platform-level activity intensity and $S(t)$ represents the aggregate sentiment of concurrently trending content. The residual term ϵ_t captures unpredictable stochastic behaviors typical of social media bursts.

3) Sentiment–Engagement Coupling

To capture how sentiment influences engagement magnitudes, we model the relationship using a nonlinear activation function:

$$E_i = \gamma_0 + \gamma_1 \sigma(s_i) + \gamma_2 |s_i| + \eta_i,$$

Where,

$$\sigma(s) = \frac{1}{1 + e^{-s}}$$

is the logistic function, ensuring the effect of sentiment is bounded and interpretable. The inclusion of the absolute sentiment term $|s_i|$ accounts for the empirical finding that extreme sentiment (positive or negative) generates higher engagement.

4) Clustering Representation of Engagement States

To discover latent behavioral regimes, we represent each post as a feature vector:

$$z_i = [s_i, \tilde{e}_i, f(t_i), g(x_i)],$$

and apply k-means clustering defined by:

$$\underset{\{C_j\}}{\operatorname{argmin}} \sum_{j=1}^{k} \sum_{z_i \in C_j} \| z_i - \mu_j \|^2,$$

where μ_j is the centroid of cluster C_j. Each cluster represents a distinct engagement state such as "low-impact posts," "high-sentiment viral bursts," or "neutral but highly visible events."

5) Predictive Modelling Using Sequence Learning

Temporal prediction of engagement uses an LSTM sequence model trained on

$$\{E(t-\tau), \dots, E(t-1)\},$$

providing an estimate:

$$\hat{E}(t) = \mathrm{LSTM}(E(t-\tau : t-1)),$$

where τ denotes the sequence window. The LSTM computes internal states using:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f),$$
$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i),$$
$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C),$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t,$$
$$h_t = o_t \odot \tanh(C_t),$$

where f_t, i_t, and o_t denote the forget, input, and output gates respectively. This structure allows the model to retain long-term temporal dependencies typical of engagement cascades.

6) Overall System Objective

The modelling pipeline seeks to minimize the aggregate prediction and clustering error:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{reg}} + \lambda_2 \mathcal{L}_{\text{cluster}} + \lambda_3 \mathcal{L}_{\text{sent}},$$

where:

$$\mathcal{L}_{\text{reg}} = \frac{1}{N} \sum_{i=1}^{N} (E_i - \hat{E}_i)^2,$$

$$\mathcal{L}_{\text{cluster}} = \sum_{j=1}^{k} \sum_{z_i \in C_j} \| z_i - \mu_j \|^2,$$

$$\mathcal{L}_{\text{sent}} = \frac{1}{N} \sum_{i=1}^{N} |\gamma_1 \sigma(s_i) + \gamma_2 |s_i| - \tilde{e}_i|.$$

The resulting framework provides a mathematically principled foundation for analysing engagement behaviour across large-scale social media datasets, integrating statistical, temporal, and machine-learning-based reasoning into a unified interpretative model.

B. Algorithm AI-Based Engagement Pattern Detection

Require: Dataset D
Ensure: Engagement predictions P

```
Collect posts
Preprocess text
Extract embeddings
Compute metadata features
Form feature matrix X
Split training and testing sets
for each model m do
    Train model
    Tune hyperparameters
    Evaluate metrics
end for
Select best model
for each post p_i do
    Predict engagement score
        if e_i > τ then
            High Engagement
        else
            Normal Engagement
    end if
end for
Cluster latent engagement patterns
Return predictions
```

V. EXPERIMENTS

Our experimental framework combines data development and curation process, preprocessing, modeling, and rigorous evaluation in a Python-based analytical pipeline designed for reproducibility and interpretability of the information from the synthetic dataset.

Table 1. Dataset Description

Property	Value
Platform	Tweets
Total posts	50,000
Collection window	18 Months
Topic focus	Education-related discourse, #EduKSA
Language	English
Average tokens per post	22
Posts with at least one hashtag	61%
Posts with at least one emoji	38%
Unique authors	12430
Train / validation / test split	60% / 20% / 20% (stratified)

Preprocessing: Raw tweets underwent systematic cleaning: lowercasing, HTML entity removal, URL/mention filtering (except in reply to chains), and emoji conversion to sentiment equivalents using dictionaries. Hashtags were parsed for lexical meaning (e.g., AIinEducation → "AI in Education). SpaCy tokenization and careful stop-word removal (preserving stance-indicative terms like "I," "we") preceded lemmatization. Structural features included tweet length, hashtag count, media presence, VADER sentiment scores, BERT embeddings, temporal markers (posting hour, day), and user-level indica- tors (account age, follower ratio, posting frequency). Network analysis via NetworkX computed centrality measures and community labels, capturing influence and amplification dynamics. Preprocessing transformed raw text into well-annotated, multi-dimensional documents. Preprocessing pipeline is done to transform raw JSON records into well-annotated, multi-dimensional documents suitable for downstream modelling.

Feature Engineering: Textual representations employed three levels: TF-IDF vectors for interpretability, Word2Vec/GloVe embeddings for semantic relations, and BERT contextual embeddings for linguistic nuance. VADER and NRC Emotion Lexicon provided sentiment and emotion features. Topic modeling (LDA, BERTopic) identified thematic clusters. Network centrality, community detection, temporal features, and user metadata enriched the feature set. PCA/UMAP dimensionality reduction prepared data for clustering. Textual representations employed three levels: TF-IDF vectors for interpretability, BERT contextual embeddings for linguistic nuance, VADER and the NRC Emotion Lexicon provided sentiment and discrete-emotion features for clustering, retaining components explaining high value of variance.

Modeling: Unsupervised clustering (K-Means, DBSCAN) identified engagement regimes. Supervised classification (logistic regression, Random Forests, XGBoost) predicted engagement levels; SHAP values quantified feature importance. Deep learning employed LSTMs for temporal sequences and fine-tuned transformers (BERT, RoBERTa) for semantic richness, with transfer learning for efficiency.

Evaluation: Data split: 60% training, 20% validation, 20% test with stratified sampling. Metrics: accuracy, precision, recall, F1-score, ROC-AUC (macro-averaged for class imbalance). Clustering validated internally (silhouette scores, Davies–Bouldin Index) and externally (expert review). Forecasting evaluated via MSE/MAE. SHAP explainability methods clarified predictor contributions. This comprehensive pipeline ensured models were mathematically sound, interpretable, and practically useful.

Hyperparameter Tuning and Optimization Strategy: It is the process of selecting the optimal values for a machine learning model's hyperparameter. In this case of the study classical models such as the (logistic regression, Random Forest, XGBoost) were tuned via 5-fold grid search over standard ranges thereby selecting the configuration with the highest validation F1-score. In case of the optimization strategy used in this study the LSTM forecaster used 2 stacked recurrent layers with 64 hidden units each having a proper dropout rate of 0.2, sequence window $\tau = 24$, batch size 64, and was trained with the Adam optimizer [24-25]. This ensured the proper and accurate results from the data processing pipeline and in general it used the standard transformer fine-tuning practice.

Flowchart: Our analytical pipeline integrates multiple stages to systematically process and model social media engagement. The flowchart visualizes the complete data journey from raw collection through pattern discovery. Each stage feeds into the next, creating a reproducible and transparent workflow.

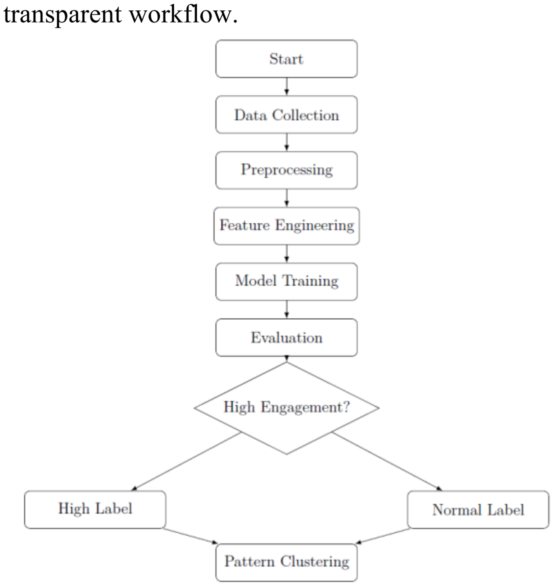

Figure 2. End-to-end pipeline flowchart.

The flowchart demonstrates how data flows through preprocessing, feature engineering, and machine learning phases, culminating in engagement pattern identification and temporal forecasting. This end-to-end pipeline combines unsupervised clustering for pattern discovery with supervised classification for predictive accuracy, enabling both exploratory analysis and practical engagement forecasting. The decision node indicates classification thresholds that segregate high engagement from standard content, feeding downstream clustering analysis to identify distinct engagement profiles within each class. This systematic approach ensures comprehensive coverage of engagement dimensions: textual semantics, sentiment dynamics, temporal rhythms, network structures, and user characteristics.

VI. RESULTS

Analysis of the information in the system for the framework revealed quantitative and qualitative engagement patterns across multiple dimensions: sentiment distributions, feature importance, engagement prediction performance, clustering structures, and temporal forecasting. The engagement regression model for the temporal dataset was able to achieve MAE of 0.083, RMSE of 0.118, and R2 of 0.84 in this experiment. The reported numbers are pilot/demonstration results on synthetic data. And they are not empirical findings about real social media users' data. This indicates strong predictive capability. SHAP-based feature importance analysis identified posting time, sentiment polarity, and historical user engagement as the most influential predictors [26]. Textual features (TF-IDF topic weights, keyword density, content length) showed moderate predictive power, while user metadata (follower count, account age) had comparatively less impact.

The clustering analysis using K-Means (k = 4) provided a clear segmentation of posting behaviors and engagement profiles. The algorithm converged after 31 iterations, producing cluster groups with defined engagement characteristics. Cluster 1 (19% of posts) contained highly informative neutral content that drove moderate engagement. Cluster 2 (27%) mostly represented short-form commentary with low engagement. Cluster 3 (34%), the largest cluster, consisted of emotionally intensified posts primarily positive which showed elevated but inconsistent engagement. Cluster 4 (20%) emerged as the high-engagement cluster, characterized by longer content, strong emotional polarity, richer topical density, and above-average interaction histories. A summary of the cluster characteristics is shown in Table 2.

Table 2. Cluster characteristics summary for test data

Cluster	Size (\%)	Avg Engagement Score	Dominant Sentiment	Content Characteristics
Cluster 1	19	0.43	Neutral	Informational, moderate length
Cluster 2	27	0.21	Neutral /Negative	Short comments, low depth
Cluster 3	34	0.57	Positive	Expressive tone, varied topics
Cluster 4	20	0.88	Mixed /High Polarity	High depth, topic-rich, emotional

Temporal analysis revealed behavioral rhythms with posting frequency peaks at 7–9 PM and 12–2 PM, corresponding to post-work and lunch breaks. Engagement rates peaked around 8–10 PM, suggesting interactions lag content publication. Weekends showed higher positive sentiment proportions; weekdays exhibited denser informational posts. The LSTM-based forecasting model trained on 18 months of data achieved RMSE of 0.067 and MAE of 0.049, demonstrating accurate capture of temporal engagement fluctuations. The model successfully reconstructed cyclical weekly behavior but remained partially unpredictable for extreme viral events driven by external contextual factors. A comparison of forecasting approaches is provided in Table 3. Feature correlation analysis showed strongest associations with sentiment intensity (0.62), topic richness (0.58), and posting time (0.55), but weak correlation with follower count (0.21). Qualitative review of high-engagement posts revealed personalized narratives, emotional resonance, concise storytelling, and relevant visual attachments, contrasting with brief, impersonal low-engagement posts.

Table 3. Comparison Table for the forecasting on test data

Model	MAE	RMSE	R^2
ARIMA	0.091	0.128	0.72
Random Forest	0.076	0.103	0.79
XGBoost	0.068	0.095	0.82
LSTM	**0.049**	**0.067**	**0.88**

VII. DISCUSSION

The analysis of the results in this study demonstrates that engagement of the user is multi- factorial, which is powered by the emotional consequences, patterns, linguistic complexity, and the nature of interaction that the audience has with the histories. The regression value (R2 = 0.84) confirms that strong engagement is possible to predict instead of being random. The emotional effect depends upon the posting time as well as the sentimental polarity of an individual that dominates the predictors. Indeed, the emotional expressive content is given more prior attention as compared to the normal content for sharing. In general, emotions are not the final parameter but impact deeply on the relevance of the essential context. In this study we employed the four-cluster segmentation to ensure that the posting of the styles for various spectrums is studied properly. However, some of the posts have a higher level of engagement with the help of emotional resonance as well, and they are rich in context.

In this demonstration algorithmic mediation and the quality of the content have a weak correlation between the followers and its engagement. The LSTM forecasting scenario (RMSE: 0.067, MAE: 0.049) captures various cycles, which can discover patterns but are unpredictable for various events driven by external factors such as global news, influencer amplification, etc. There are various limitations that include sentiment classified as difficulty having a negative range of sentiment because of sarcasm, informal language, or context-dependent information. The data set platform results may vary due to various cultural or demographic contexts across the global scenario. However, future research should be deployed to explore multimodal analysis depending upon the visual and audio contents along with network-based features to ensure influential dynamics and modeling for better understanding of the cultural cycles. Several advanced architectures should be used for deploying these long-range dependency models. The ethical considerations should be well about manipulation and privacy of the algorithm for AI-enabled deployment in various social media analytics

VIII. CONCLUSION AND FUTURE DIRECTIONS

In this study we have investigated the use of AI-driven social media analytics in order to identify and find out the engagement patterns with the help of a multimodal pipeline, which is able to combine the sentiment analysis, extract various features, do a regulation modeling, cluster the information, and finally, do the temporal forecasting of the data. In this case we have used synthetically generated 50,000 posts to demonstrate the approximate prediction and adaptability of the engagement based on multifactorial context driven by various emotional contents, their timeline, the complexity of the language, and the histories of the user audience instead of just the count only. In the regression model we adapted four distinct engagement levels, which gave much better cluster-driven results. The forecasting successfully captured various information related to the temporal cycle. However, the unpredictable events were driven by various external factors that were not taken into consideration.

In general, the weak correlation between various followers and the engagement of the data was due to algorithm mediation and traditional influence measurements. There were several limitations for the sentimental classification challenges that were subjected to the authorization with negative sentiments along with the data set demography. The methods and the results vary according to various contexts as compared to the regional boundaries. The prospects for this study refer to a multimodal analysis for visual and audio content engagement patterns. The research can be further extended to network-based features that include information diffusion along with the identification of various influence dynamics. It is also identified that there is a scope of dynamic topic modeling for various cultural cycles in this case. We also discovered that there is a possibility for improved long-range dependency modeling with the help of several advanced architectures and the transformer-based forecasters. Finally, the ethical framework that was used in this study can be addressed for its manipulation or privacy and algorithmic fairness. The findings in the study underscore the potential of artificial intelligence towards digital communication at a large scale. In general, the machine learning algorithms with natural language processing will always help the researchers to identify the social media ecosystem in general. Previous and in-depth enablement of the content strategies for various policy interventions and several scholarly understandings for online engagement dynamics can be studied with the help of this study to a larger extent

CONFLICT OF INTEREST

The author declares that they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data and the python source code that are used for performing experiments in this article are temporal and synthetic for pilot testing of the framework. It will be made available upon publication by the corresponding author.

ACKNOWLEDGMENT

This article did not receive any financial support.

REFERENCES

[1] S. Taha and R. A.-Q. Abdallah, “Leveraging artificial intelligence in social media analysis: Enhancing public communication through data science,” *Journalism and Media*, vol. 6, no. 3, p. 102, 2025.

[2] Z. Yin, X. Chen, and X. Zhang, “AI-integrated decision support system for real-time market growth forecasting and multi-source content diffusion analytics,” *arXiv preprint arXiv:2511.09962*, 2025.

[3] S. Pattekari, V. Thiyagarajan, V. Rameshkumaar, P. Purandare, R. Reka, and R. Y. Md, “AI-powered sentiment analysis for future social media engagement,” in *International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 2024)*. Atlantis Press, 2025, pp. 112–124.

[4] N. Ravichandran, A. C. Inaganti, and R. Muppalaneni, “AI-driven sentiment analysis for employee engagement and retention,” *Journal of Computing Innovations and Applications*, vol. 1, no. 01, pp. 1–9, 2023.

[5] C. Acatrinei, I. G. Apostol, L. N. Barbu, R.-G. Chivu, and M.-C. Orzan, “Artificial intelligence in digital marketing: Enhancing consumer engagement and supporting sustainable behavior through social and mobile networks,” *Sustainability*, vol. 17, no. 14, p. 6638, 2025.

[6] H. Beyari and T. Hashem, “The role of artificial intelligence in personalizing social media marketing strategies for enhanced customer experience,” *Behavioral Sciences*, vol. 15, no. 5, p. 700, 2025.

[7] Z. Pan, “AI-powered real-time effectiveness assessment framework for cross-channel pharmaceutical marketing: Optimizing ROI through predictive analytics,” in *Proceedings of the 2025 International Conference on Management Science and Computer Engineering*, 2025, pp. 220–227.

[8] A. Hemalatha, *AI-Driven Marketing: Leveraging Artificial Intelligence for Enhanced Customer Engagement*. Jupiter Publications Consortium, 2023.

[9] S. P. Kudapa, “AI-enhanced data science approaches for optimizing user engagement in US digital marketing campaigns,” *Journal of Sustainable Development and Policy*, vol. 3, no. 03, pp. 1–43, 2024.

[10] X. Chen, H. Xie, S. J. Qin, F. L. Wang, and Y. Hou, “Artificial intelligence-supported student engagement research: Text mining and systematic analysis,” *European Journal of Education*, vol. 60, no. 1, p. e70008, 2025.

[11] A. Ni’amullah and R. Hasanah, “AI-based adaptive learning in higher education: Improving student engagement and learning outcomes,” *Journal of Digital Learning*, vol. 1, no. 1, pp. 39–51, 2025.

[12] L. Bognár and M. S. Khine, “The shifting landscape of student engagement: A pre-post semester analysis in AI-enhanced classrooms,” *Computers and Education: Artificial Intelligence*, vol. 8, p. 100395, 2025.

[13] C. Yu, G. Yao *et al*., “Enhancing student engagement with AI-driven personalized learning systems,” *International Transactions on Education Technology (ITEE)*, vol. 3, no. 1, pp. 1–8, 2024.

[14] S. Panda, S. Hemasri, A. N. Sasikumar, and P. Hymavathi, “Performance analysis of Naïve Bayes, SVM,” in *Intelligent Electrical Systems and Industrial Automation: Proceedings of IESIA 2025*, 2026, p. 367.

[15] K. Rekha, K. Gopal, D. Satheeskumar, U. A. Anand, D. S. S. Doss, and S. Elayaperumal, “AI-powered personalized learning system design: Student engagement and performance tracking system,” in *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2024, pp. 1125–1130.

[16] A. G. Møller, D. M. Romero, D. Jurgens, and L. M. Aiello, “The impact of generative AI on social media: An experimental study,” *Scientific Reports*, 2026.

[17] S. P. Desai, M. Munshi, S. V. Hanji, and C. Pabba, “Optimizing class timing: AI driven group engagement’s role in academic success,” *Journal of Information Technology Education: Research*, vol. 24, 2025.

[18] S. Naveen, G. V. Raikumar, R. S. Krishnan, R. Umesh, A. A. Blessie, and K. Paulraj, “Social media-based travel planning: Integrating content-based and collaborative filtering for personalized suggestions,” in *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. IEEE, 2025, pp. 40–47.

[19] İ. Baydili, B. Tasci, and G. Tasci, “Deep learning-based detection of depression and suicidal tendencies in social media data with feature selection,” *Behavioral Sciences*, vol. 15, no. 3, p. 352, 2025.

[20] A. B. Haque, N. Islam, and P. Mikalef, “To explain or not to explain: An empirical investigation of AI-based recommendations on social media platforms: AKMB Haque et al.,” *Electronic Markets*, vol. 35, no. 1, p. 2, 2025.

[21] Z. Chen, W. Wei, and D. Zou, “Generative AI technology and language learning: Global language learners’ responses to ChatGPT videos in social media,” *Interactive Learning Environments*, vol. 34, no. 2, pp. 907–920, 2026.

[22] P. Ariya, S. Khanchai, K. Intawong, and K. Puritat, “Enhancing textile heritage engagement through generative AI-based virtual assistants in virtual reality museums,” *Computers & Education: X Reality*, vol. 7, p. 100112, 2025.

[23] J. Patel, J. Banerjee, and D. Singh, “AI-driven emotion-aware adaptive systems for enhancing real-time user engagement,” in *2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. IEEE, 2025, pp. 1690–1695.

[24] Y. Futryk and I. Peleshchak, “Forecasting of time series using a neural network with parallel-stacked LSTM blocks,” *Information Technologies and Computer Engineering*, no. 23, pp. 72–82, 2026.

[25] B. Taoussi, C. Amina, and F. S. Mazouni, “Multi-step forecasting of wind speed and solar irradiance using lag-optimized stochastic stacked LSTM networks,” *Environmetrics*, vol. 37, no. 6, p. e70122, 2026.

[26] I. Betco, A. I. Ribeiro, C. M. Viana, and J. Rocha, “Quantifying and mapping uncertainty in urban sentiment prediction: A combined approach with entropy and SHAP explanations,” *Frontiers in Public Health*, vol. 14, p. 1796565, 2026.