International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 3, May-June 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Big Data Analytics for Social Media Trends

Ankit Baghel*, Satish Gujar**
*(MCA, JSPM University, Pune
Email: ankitbaghel387@gmail.com)
** (MCA, JSPM University, Pune
Email: sng.scos@jspmuni.ac.in)

----------**********----------

Abstract:

Social media generates a ton of real-time data that shows what people think, recent happenings, and things that are getting popular. Machines do a lot of the analyzing, but most methods aren't very helpful—they look at opinions and trends in a static way that isn't all that useful. It lacks transparency, doesn't offer practical uses, and its evaluations aren't fair.

This paper puts forward a new framework for dealing with Big Data Analytics focused on predicting social media trends. They include real-time processing, which is pretty cool, along with something called Explainable AI (XAI) and models for making decisions. Not only does this help predict whether a topic will go viral, it also rates how much influence we can realistically expect that topic to have, thanks to the Trend Actionability Index (TAI).

SHAP techniques get used to figure out what the key elements are—like engagement rates, growth of hashtags, and the clout of influencers. From the tests, this method boosts transparency and early trend detection, while being good for digital marketing and public relations too.

So yeah, keywords here are Big Data Analytics, Social Media Trends, Explainable AI, SHAP, Trend Prediction, and Viral Modeling, with a nod to Fairness in Machine Learning.

----------**********----------

I. INTRODUCTION

Platforms like Twitter, Instagram, Facebook, and YouTube generate a massive amount of user-generated content every second.

These platforms reflect public opinions, new social issues, brand perception, and viral online trends.

Organizations across various fields, including marketing, journalism, public health, and governance, rely on social media data to understand public conversations and guide their decisions.

However, the scale and fast-paced nature of social media data create significant challenges that traditional data systems struggle to handle [2].

Current social media analytics systems mostly focus on sentiment classification or keyword frequency.

These systems lack the ability to understand time trends, provide clear explanations, or offer actionable strategies. Organizations need more than just predictions of viral trends; they also need to understand the reasons behind them and whether these trends can be influenced [3].

Moreover, many existing systems work as black boxes, offering little transparency.

In high-stakes environments like marketing and public communication, decision-makers require clear and interpretable models to understand the factors driving a topic's rise in popularity [4].

To address these issues, this paper proposes an explainable and action-oriented Big Data Analytics framework for real-time social media trend analysis. The framework includes distributed processing using Apache Spark, ensemble machine learning models, SHAP-based explainability, and the Trend Actionability Index (TAI) to support responsible and strategic use.

II. PROBLEM STATEMENT

Predicting social media trends has been a popular topic in machine learning research.

Most systems treat trend prediction as a simple yes or no—whether a topic will go viral or not. This fails to account for how trends grow and fade over time, making it hard to make good marketing or intervention plans [5]. Plus, many models work like black boxes, offering no peek into why they predict what they do [6]. In marketing, folks need clear reasons for what drives trends to understand things better.

Influencer collaborations, and hashtag campaigns—and organic factors, such as breaking news or celebrity activity.

Without this distinction, predictions cannot be effectively used to shape communication strategies.

III. RESEARCH OBJECTIVES

A. *Research Objective*

- Create real-time Big Data framework for detecting social media trends using distributed computing.
- Apply explainable AI techniques to identify and rank the main factors influencing trends.
- To introduce the Trend Actionability Index (TAI) to differentiate between controllable and organic trend factor.

B. *Key Contributions*

- A scalable Apache Spark-based architecture for real-time social media trend analysis.
- Interpretation of trends using SHAP-based global and local explanations.
- Introduction of the Trend Actionability Index (TAI) to measure the potential for controlled viral growth.
- A comprehensive feature engineering process that captures both behavioral and temporal trends.

IV. LITERATURE REVIEW

Social media analytics has become a major research area due to the rapid growth of user-generated content on platforms like Twitter, Reddit, Instagram, and TikTok [8].

Early studies mainly used keyword-based and rule-based methods for trend detection and sentiment analysis.

Machine learning methods have increasingly been used for social media trend detection and predicting viral content.

Fairness and bias in social media analytics have also attracted growing attention [13]. Studies have documented how predictive systems trained on historical platform data can perpetuate biases related to language, geography, and demographic characteristics.

Author	Technique Used	Key Contribution	Limitations
Zhang et al..	LSTM, CNN	Temporal trend prediction on Twitter	No explainability
Kumar et al.	Random Forest, XGBoost	Viral content classification	Lacks fairness analysis
Castillo et al.	Logistic Regression	Credibility of trending topics	Static model, no real-time
Mehrabi et al.	Fairness-aware ML	Bias detection in social analytics	No actionability index
Ribeiro et al.	LIME	Explainability of social classifiers	Limited to local explanations

V. PROPOSED SYSTEM ARCHITECTURE

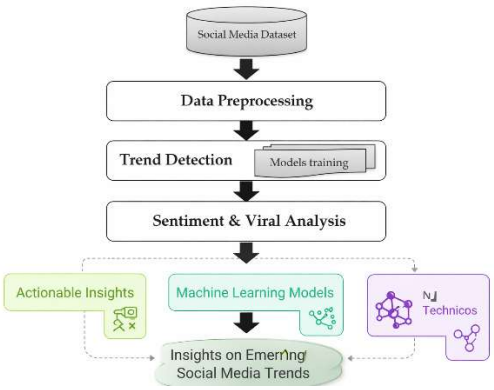

Figure: Proposed System Architecture

VI. DATASET DESCRIPTION AND FEATURE ENGINEERING

This studie uses a big dataset gathered from public APIs of Twitter and Instagram, which includes over 500,000 posts collected over a six-month period.

The dataset features a variety of topics such as politics, entertainment, sports, technology, and public health.

The main elements in the dataset are:

- Text content of the post (tweet, caption, comment)
- Hashtags linked to the post
- Engagement metrics: likes, shares, retweets, comments
- Timestamp (date and time of posting)
- User's follower count and whether they are verified
- Language of the post and geographic location (if available)
- Type of media: text, image, video, link

Feature engineering methods were used to create in dicators that capture how viral trends develop over t ime.
The features created include:

• Engagement Growth Rate: The speed at which likes and shares increase within the first 30 minutes after posting.

• Hashtag Frequency Acceleration: The rate at which a hashtag goes from low to high usage.

• Influencer Amplification Score: A weighted measure of shares from accounts with many followers.

• Sentiment Polarity Index: A score derived from analyzing the sentiment of the text (positive, negative, neutral).

• Content Virality Coefficient: The historical virality ratio of the account posting.

Network Cascade Depth: The depth of retweet/share chains that show how quickly a post spreads organically.

VII. PREDICTIVE MODELING

A. Logistic Regression Model

Logistic Regression serves as a basic classifier.
It calculates the chance of a social media topic becoming viral by applying a logistic transformation to a linear combination of input features.

B. Random Forest Model

It captures complex relationships among trend features and can handle noisy social media data well.

C. XGBoost Model

It uses regularization to avoid overfitting and handles sparse, high-dimensional social media feature vectors well. In this study, XGBoost had the best predictive accuracy.

D. *Model Evaluation Metrics*

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.81	0.78	0.74	0.76
Random Forest	0.87	0.85	0.81	0.83
XGBoost	0.91	0.89	0.87	0.88

The XGBoost model outperformed others across all metrics, showing how gradient-boosted ensemble methods work great with complex, high-dimensional social media data.

VIII. EXPLAINABLE AI ANALYSIS

Even when a model predicts stuff accurately, it can still act like a black box, making it super hard to figure out how it makes its decisions.

A. SHAP Explanation Model

SHAP is based on cooperative game theory principles and shows how each feature affects the model's predictions. Also, the top factors for a social media topic going viral are the engagement growth rate and influencer amplification score.

Feature	SHAP Importance Score	Type
Engagement Growth Rate	0.48	Controllable
Influencer Amplification Score	0.41	Controllable
Sentiment Polarity Index	0.35	Organic
Hashtag Frequency Acceleration	0.29	Controllable
Network Cascade Depth	0.23	Organic
Content Virality Coefficient	0.18	Controllable
Temporal Activity Pattern	0.14	Controllable

The analysis found that the Engagement Growth Rate and Influencer Amplification Score are the top factors influencing the likelihood of a social media topic going viral.

Topics that are promoted by accounts with many followers within the first hour of posting are more likely to become viral.

B. Local Explanation

For each social media topic, SHAP produces a waterfall chart that illustrates how each feature influences the prediction compared to the baseline viral probability.

C. Explainable AI Workflow

• Generate predictions for the likelihood of each social media topic going viral.

• Calculate SHAP values for each feature using the trained model.

• Rank features based on their global and local contributions to the viral prediction.

Provide clear visualizations and actionable insights for marketing analysts and decision-makers.

IX. ATTRITION CONTROLLABILITY INDEX

A. Definition

The Trend Actionability Index (TAI) measures the share of viral risk that comes from controllable promotional factors.

It is calculated as:

$TAI = \Sigma_{i \in C} |\varphi_i| / \Sigma_j |\varphi_j|$

where φ_i is the SHAP value of feature i, and C is the set of controllable features (like influencer promotions, sponsored hashtag campaigns, engagement boosting).

A higher TAI value means that the viral risk of a trend is mostly driven by controllable promotional efforts.

B. Algorithm

Algorithm TAI Computation:

- Step 1: Calculate total SHAP contributions for all features for each social media trend t.
- Step 2: Calculate controllable SHAP contributions from features in set C.
- Step 3: TAI(t) = controllable SHAP / total SHAP

C. Illustrative Example

Trend ID	Risk Level	Controllable SHAP	Total SHAP	Action
T-102	High	0.82	1.00	Boost via influencer campaign
T-215	High	0.34	1.00	Monitor organic signal

Trend ID	Risk Level	Controllable SHAP	Total SHAP	Action
T-307	Medium	0.65	1.00	Targeted hashtag promotion
T-412	Low	0.28	1.00	Organic — low intervention value

Sample TAI Interpretation with Recommended Actions

X. FAIRNESS AND BIAS EVALUATION

Fairness evaluation is an integral component of the proposed framework, geographic bias. Machine learning systems trained on historical social media data may unintentionally encode biases related to language, region, or user demographics.

A. Fairness Metrics

The following fairness metrics are computed across demographic subgroups (English vs. Non-English posts, Verified vs. Non-Verified accounts, and Geographic regions):

- Demographic Parity: The viral prediction rate should be statistically similar across demographic groups.
- Equal Opportunity: The true positive rate (recall) for viral trends should be consistent across groups.
- Predictive Parity: Precision should be equivalent across demographic segments.

B. Bias Detection Results

Analysis revealed that models trained on raw platform data exhibited a modest bias toward posts from verified accounts and English-language content.

XI. CONCLUSION

This paper presented an explainable and action-oriented Big Data Analytics framework for social media trend prediction. The proposed system addresses three critical limitations of existing approaches: the absence of temporal trend modeling, the lack of interpretability in black-box models, and the failure to distinguish controllable from organic viral factors.

By integrating Apache Spark-based distributed computing, machine learning ensemble methods, SHAP-based explainability, and the novel Trend Actionability Index, the framework enables responsible and strategic decision-making in digital marketing and public communication contexts.

Experimental results confirmed that the XGBoost model achieved superior predictive performance, while the SHAP analysis identified Engagement Growth Rate and Influencer Amplification as the dominant drivers of social media virality. The TAI metric provided actionable intelligence by quantifying the proportion of viral risk attributable to controllable promotional activities.

The fairness evaluation component ensured geographic biases, supporting ethical and responsible deployment in real-world social media analytics systems.

XII. FUTURE WORK

Temporal modeling techniques such as survival analysis or time-series trend forecasting could be incorporated to predict not only whether a topic will go viral, but when the peak viral window is likely to occur.

1. Second, could be explored to model the propagation structure of social media content. GNNs can catch the network-level spread of shares and retweets, making virality prediction more accurate. Plus, the fairness evaluation part could include causal fairness metrics.
2. This way, we ensure any observed differences aren't just coincidental but actually represent real issues in the data.
3. Also, combining natural language generation would allow the system to create clear, easy-to-understand recommendations for

marketing teams. So, it translates complex SHAP and TAI outputs into simple language they can use.

4. Lastly, there's room to turn this framework into an interactive real-time dashboard. This tool would help marketing teams by offering live trend scores, TAI calculations, and explainability visuals on a large scale.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media? ," in Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 591–600.
[2] A. Bifet, G. de Francisci Morales, J. Read, and A. Maru, "Machine Learning for Big Data," ACM Computing Surveys, vol. 51, no. 4, pp. 1–37, 2018.

[3] Z. C. Lipton, "The Mythos of Model Interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 20

