(IJRST) 2024, Vol. No. 14, Issue No. 4, Oct-Dec **e-ISSN: 2249-0604, p-ISSN: 2454-180X**

Developing a Machine-Learning Based Smart Integrated Model Trained on 50M Instagram Reels Meta-Data to Enhance the Predictability of Viral Moments

Om Venkatesh Sharma

Delhi Public School, Vasant Kunj

DOI:10.37648/ijrst.v14i04.008

[1]*Received: 08 October 2024; Accepted: 19 November 2024; Published: 29 November 2024*

ABSTRACT

In today's fast-paced digital landscape, short-form video content has emerged as a dominant mode of user engagement, particularly through platforms like Instagram Reels. As creators, marketers, and platforms seek to understand what makes content go viral, predictive analytics powered by large-scale data has become a compelling research frontier. This paper presents a machine-learning-based approach to predicting viral moments using metadata from 50 million publicly available Instagram Reels. We define virality using statistical thresholds on engagement metrics such as views, likes, shares, and comments, identifying the top 10% as "viral."

Our methodology involves comprehensive feature extraction, including hashtag count, audio trends, post timing, caption sentiment, and user interaction levels. We implemented and compared three supervised models: logistic regression, random forest, and XGBoost. Among these, XGBoost delivered the highest performance with 87% accuracy and an AUC of 0.88, demonstrating strong capability in capturing nonlinear relationships and interaction effects among features.

Key findings show that the use of trending audio, optimal posting hours (evenings and weekends), and higher hashtag density significantly correlate with virality. The model also highlights the predictive value of caption sentiment and engagement velocity within the first hour of posting.

This research not only offers a practical tool for content creators and marketers but also contributes to the growing literature on algorithmic content optimization and social media analytics. Future work can enhance prediction capabilities by integrating visual and audio features from the actual video content and adapting the model for real-time deployment on social platforms.

1. Introduction

The explosion of short-form video content has fundamentally transformed the way individuals consume and interact with media. Among these formats, Instagram Reels has emerged as a powerhouse platform, engaging billions of users worldwide and redefining the dynamics of online content creation. With Instagram's growing focus on video, Reels have become one of the most influential mediums for creators, brands, influencers, and everyday users to communicate and entertain. The increasing prominence of these videos has led to a pressing need to understand and predict what makes a Reel go "viral."

Virality in digital content is defined as the rapid and widespread sharing of media across a network, often fueled by social signals like shares, likes, comments, and algorithmic boosts. For content creators and marketers, the ability to predict virality is more than a curiosity—it is a strategic necessity. Platforms like Instagram increasingly rely on

[1] How to cite the article: Sharma O.V; December 2024; Developing a Machine-Learning Based Smart Integrated Model Trained on 50M Instagram Reels Meta-Data to Enhance the Predictability of Viral Moments; *International Journal of Research in Science and Technology*, Vol 14, Issue 3, 70-74, DOI: http://doi.org/10.37648/ijrst.v14i04.008

machine-learning-based recommendation engines to surface content that is more likely to engage users. Hence, being able to reverse-engineer or anticipate this algorithmic preference becomes a valuable tool.

Despite significant interest in virality prediction, the research space remains fragmented, with prior studies often focusing on text-based social media (e.g., Twitter) or long-form video platforms like YouTube. These models typically emphasize content semantics, network graphs, or user influence. However, Instagram Reels present unique challenges and opportunities: they are short, visually rich, driven by audio trends, and highly reactive to timing and interaction dynamics.

This study fills a critical gap by exploring the predictive power of metadata associated with Reels, rather than the content itself. Metadata such as post time, duration, audio usage, hashtag count, caption length, and early engagement statistics offer a scalable and ethical way to evaluate potential virality. Analyzing metadata also avoids the privacy and storage complications associated with processing millions of videos directly.

To address this, we curated a dataset of over 50 million publicly available Instagram Reels across diverse content categories and engagement levels. We defined virality statistically using the 90th percentile threshold of combined views and shares. Features were extracted, engineered, and normalized to create a rich input space for training supervised learning models.

We compare three classification models—logistic regression (as a baseline), random forest (to capture nonlinear dependencies), and XGBoost (an advanced ensemble method known for high performance in classification problems). The models were evaluated using metrics such as accuracy, F1-score, and area under the ROC curve (AUC), allowing us to assess not just predictive power but also generalization capability.

Our findings reveal important insights into what drives content virality on Reels. Factors such as trending audio, optimized posting time (evenings and weekends), hashtag density, and early interaction velocity are shown to be key predictors. These insights can help content creators optimize their posting strategies, guide marketing campaigns, and improve platform-level content recommendations.

In summary, this research contributes to the field of computational social science and predictive analytics by offering a robust, scalable framework for virality prediction using only metadata—an approach that balances performance, efficiency, and ethical concerns in today's data-sensitive ecosystem.

Table 1: Commonly Observed Viral Metrics in Literature

Metric	Description	Common Threshold
Views	Total number of times a Reel is played	>100,000
Likes	Number of user likes	>10,000
Shares	Number of times the Reel is shared	>5,000
Comments	Total user comments	>500

2. Methodology

We collected metadata from 50 million public Instagram Reels over a 12-month period using ethical web-scraping and API-based tools. The metadata included: (1) View Count, (2) Like Count, (3) Hashtag Count, (4) Audio Usage (original or trending), (5) Post Timing (hour and day), (6) Captions, (7) Comments, and (8) Duration. We labeled a Reel as "viral" if it surpassed the 90th percentile in both views and shares.

Feature engineering involved one-hot encoding for categorical variables (e.g., day of week), normalization for continuous variables, and TF-IDF for caption text. Three models were implemented: Logistic Regression (baseline), Random Forest (nonlinear), and XGBoost (boosted ensemble).

Table 2: Feature List and Encoding Strategy

Feature	Encoding/Transformation
Hashtag Count	Numerical (normalized)
Audio Type	One-hot encoding
Post Time (hour)	Binned categories
Caption Text	TF-IDF vectorization
Likes, Comments	Log transformation

3. Results

This section presents the evaluation outcomes of the three machine-learning models—Logistic Regression, Random Forest, and XGBoost—applied to the Instagram Reels metadata for virality prediction. The models were trained using a stratified 80/20 train-test split to ensure balanced representation of viral and non-viral Reels. Performance metrics such as accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC) were used to assess model effectiveness.

The baseline Logistic Regression model yielded moderate performance with an accuracy of **72.4%**, F1-score of **0.69**, and AUC of **0.73**. Although efficient in terms of training time and interpretability, it failed to capture the complex, nonlinear relationships inherent in content virality. Variables like audio trends, interaction effects between hashtags and post timing, and caption sentiment were only partially modeled through linear relationships, leading to underfitting in many cases.

The Random Forest model improved performance across all metrics, achieving an accuracy of **81.2%**, F1-score of **0.78**, and AUC of **0.81**. The ensemble method successfully captured some nonlinear interactions and benefited from decision tree diversity. However, the model demonstrated some degree of overfitting, particularly in minority classes (i.e., borderline viral posts), which impacted generalization to unseen data. The interpretability of individual decisions was also reduced due to the ensemble nature of the model.

The XGBoost model delivered the highest performance, achieving **87.0%** accuracy, **0.85** F1-score, and **0.88** AUC. It excelled in capturing nuanced relationships between features and their composite impact on virality. For example, Reels using trending audio and posted between 7 PM and 10 PM with 5+ hashtags were significantly more likely to be classified as viral. Additionally, Reels with high interaction velocity (likes/comments within the first 30 minutes) were robust predictors of future viral status. The model also showed stability across multiple test folds, indicating high reliability.

The confusion matrix for the XGBoost model indicated high precision and recall for both viral and non-viral classes. False positives were mostly Reels with high view counts but lower shares, suggesting that shares are a better virality indicator than views alone. Feature importance analysis ranked “Audio Usage” as the top predictor, followed by “Hashtag Count,” “Posting Time,” and “Caption Sentiment.” Interestingly, caption length had a negative correlation with virality beyond a certain word count, suggesting that overly long captions may dilute engagement.

Table 3: Model Performance Metrics

Model	Accuracy	F1-Score	AUC
Logistic Regression	72.4%	0.69	0.73
Random Forest	81.2%	0.78	0.81
XGBoost	87.0%	0.85	0.88

4. Discussion

Our study shows that viral prediction is inherently nonlinear, relying on interactions between multiple features such as sound usage, timing, and user engagement. Trending audio was associated with a 26% increase in virality likelihood. Weekend posting yielded higher engagement rates. The XGBoost model's ability to capture feature interactions was critical. These findings can aid influencers and brands in scheduling content and enhancing creative strategies.

Table 4: Feature Importance in XGBoost

Feature	Importance Score
Audio Usage	0.27
Hashtag Count	0.21
Posting Time	0.18
Caption Sentiment	0.15
Duration	0.11

5. Conclusion

This study set out to explore the feasibility and effectiveness of predicting Instagram Reel virality using only metadata from a vast dataset comprising 50 million public posts. In doing so, it establishes a scalable, interpretable, and ethically viable model for identifying the factors that drive engagement in short-form video content. Through a comparative analysis of logistic regression, random forest, and XGBoost classifiers, we demonstrate that metadata features alone—without any image, video, or audio content—can offer significant predictive power.

Among the models tested, XGBoost consistently outperformed the others across all evaluation metrics, achieving an accuracy of 87%, an F1-score of 0.85, and an AUC of 0.88. These results confirm that nonlinear relationships between metadata features—such as the interaction between trending audio and post timing—are critical to understanding virality. Notably, features like trending sound usage, hashtag count, early interaction rates, and optimized posting times were identified as strong predictors. These insights provide a valuable roadmap for content creators and social media marketers aiming to increase their content reach strategically.

Importantly, this study underscores the potential of machine learning in enhancing content strategy without infringing on user privacy, as it relies solely on publicly available and non-intrusive data. Future work may extend this approach by incorporating multimodal inputs such as frame-level video features, sound characteristics, and deep-learning-based sentiment analysis.

In conclusion, the research not only advances the state of viral content prediction but also provides practical implications for designing data-driven, ethical strategies for digital influence and engagement.

6. References

Borge-Holthoefer, J., Baños, R. A., Gonzalez-Bailon, S., & Moreno, Y. (2013). Cascading behavior in complex socio-technical networks. *Journal of Complex Networks, 1*(1), 3–24. https://doi.org/10.1093/comnet/cnt006

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? *Proceedings of the 23rd International Conference on World Wide Web*, 925–936. https://doi.org/10.1145/2566486.2567997

Kemp, S. (2023). *Digital 2023: Global overview report*. DataReportal. https://doi.org/10.5281/zenodo.7619407

Yang, J., & Leskovec, J. (2010). Modeling information diffusion in implicit networks. *2010 IEEE International Conference on Data Mining*, 599–608. https://doi.org/10.1109/ICDM.2010.74