ISSN: **3107-6513**

Instagram Reach Analysis Using Machine Learning

Vijay Shankar[#1], Utkarsh Sharma[*2], Ayush Kumar[#3]

[#1]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[*2]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[#3]*Student, Department of Information Science & Engineering, CMR Institute of Technology, Bengaluru-560037, Karnataka, India*

[1]vish22ise@cmrit.ac.in, [2]utsha22ise@cmrit.ac.in, [3]ayuku22ise@cmrit.ac.in

***Abstract*—Instagram is a major platform for digital communication and brand promotion, yet achieving consistent post reach remains challenging due to limited analytical insights. This study proposes a data-driven machine learning approach to predict Instagram reach by analysing early engagement signals such as likes, comments, saves, shares, and profile visits. Multiple regression-based models are evaluated using standard performance metrics, and feature importance analysis is conducted to identify key factors influencing reach. The results highlight that combined engagement indicators significantly impact content visibility, enabling more effective and predictive Instagram content strategies.**

***Keywords*—Instagram Reach Analysis, Social Media Analyt ics, Machine Learning, Engagement Metrics, Linear Regression**

I. INTRODUCTION

Th Social media platforms have become an integral part of communication, marketing, and digital interaction in the mod ern era. Among these platforms, Instagram has emerged as one of the most influential spaces for individuals, content creators, and businesses seeking to build visibility and engage with audiences [1]. While metrics such as likes and impressions provide surface level feedback, reach which is the number of unique users who view a post remains the most critical indicator of content visibility. Optimizing reach is essential for achieving sustained audience growth and effective brand presence [2]. However, despite the availability of native analytics tools, Instagram continues to function as a black box for many users. Creators often find it difficult to understand why certain posts achieve high reach while others fail to gain attraction. This uncertainty arises from the complex interaction of multiple factors such as posting time, hashtags, media format, caption content, and early audience engagement [3]. The non-linear and dynamic nature of these interactions makes intuitive or manual analysis insufficient for consistent performance improvement. Machine learning has become a powerful tool for addressing such complexity in social media analytics. Social platforms generate huge amount of engage ment data on a continuous basis, making the traditional rule based or statistical analysis techniques inefficient. Machine learning models are capable of automatically learning patterns from historical data and capturing hidden relationships among multiple engagement metrics [4].

Instagram data is particularly dynamic and multi-dimensional, involving features such as likes, comments, shares, saves, impressions, follower count, and profile visits, all of which interact in non-linear ways. Machine learning models are well suited to this environment because they can adapt to evolving user behavior and algorithmic changes over time. Unlike static analytical approaches [5], these models can be retrained with new data, ensuring that predictions remain relevant even as platform dynamics shift. Another key advantage of ma chine learning lies in its ability to provide interpretability [6] alongside prediction. Modern regression and ensemble learning techniques can highlight the relative importance of different engagement features, helping users understand not only what content performs well, but why it performs well. This interpretability transforms social media analytics from descriptive reporting into predictive and prescriptive intelligence. In the context of Instagram reach analysis, early engagement signals such as likes, comments, saves, and shares act as indicators of content relevance and audience interest. By learning from the historical post data, machine learning models can estimate the potential reach of new posts and identify high-impact engagement drivers [7], enabling more informed content planning.

Despite these advancements, existing Instagram analytics systems remain largely descriptive in nature [8]. Most creators and businesses rely on the built-in Instagram Insights inter face or third-party dashboards that primarily report historical metrics such as impressions, reach, and engagement counts. While these tools are effective for data aggregation, they fail to explain the underlying reasons behind reach fluctuations. Engagement metrics are presented as isolated values without analyzing their interdependence or combined influence on visibility [9]. Furthermore, current systems lack predictive capabilities and do not offer actionable insights for optimizing future posts. As a result, users are often forced to rely on trial & error strategies, leading to inconsistent growth and inefficient content planning. The traditional approach to Instagram analytics also suffers from several limitations [7]. Metrics are reported passively, without contextual interpretation or forecasting support. Important insights such as how a specific combination of posting time, hashtag usage, and caption length influences reach are rarely provided. Audience demographics are often presented in broad categories, making it difficult to target interests or behavioral patterns. These shortcomings highlight the need for a more intelligent, data driven framework that can move beyond surface level observation in addition to system limitations, social media data mining presents inherent challenges. Instagram data is highly volatile, with trends and algorithmic preferences changing rapidly. Content elements that perform well at one point in

time may lose relevance due to moderation policies or algorithm updates. Data access is also constrained by privacy regulations and API rate limits, which restrict the volume and granularity of available data. Moreover, social media interactions are strongly influenced by human behavior, including emotional appeal, visual aesthetics, cultural context, and timing factors that are difficult to quantify directly. These subjective elements often require proxy metrics such as saves and shares to approximate content value and relevance.

II. Literature Survey

Existing research on Instagram analytics and social media performance has explored a wide range of perspectives, including growth strategies, influencer dynamics, engagement modelling, content analysis, and machine learning based prediction [10]. Collectively, these studies highlight that Instagram reach and engagement are influenced by multiple interacting factors rather than a single metric, reinforcing the need for data-driven analytical approaches. Several studies focus on growth strategies and algorithmic visibility. Bellavista [11] examines automated interaction techniques such as follow unfollow strategies, targeted engagement, and hashtag-based user discovery, demonstrating that consistent interaction significantly increases reach and follower growth. While their work shows how automation can amplify visibility, it primarily emphasizes action driven growth rather than understanding the intrinsic relationships between engagement metrics and reach.

This distinction highlights a gap between operational growth tactics and analytical reach modelling. Another major research direction emphasizes content characteristics and influencer identification. Alwan [12] proposes a content driven approach for identifying influential Instagram users by analyzing visual features of posts using deep learning models. Their findings show that image composition, visual clarity, and aesthetic quality strongly correlate with influence, sometimes outperforming traditional metrics such as follower count. Although this work does not explicitly model reach, it reinforces the importance of content quality as a latent driver of engagement and visibility [13]. Studies focusing directly on reach analysis provide more targeted insights. Nemade [14] conducted a comparative analysis of Instagram content formats and demonstrate that Reels consistently achieve higher reach than posts, stories, or IGTV. Their work highlights the role of posting time, hashtag usage, and format selection in maximizing visibility. However, the analysis is largely descriptive and visualization-based, lacking predictive modelling. Similarly, Srushti [15] approach reach prediction as a time series forecasting problem using SARIMA models, showing that temporal patterns and seasonality can be leveraged to estimate future reach. While effective for historical trend analysis, such approaches are limited in capturing engagement driven dynamics and sudden viral spikes [16]. Machine learning based predictive studies form a core foundation for this project.

Singh applied supervised regression algorithms to predict engagement metrics using features such as follower count, hashtags, captions, and posting time. Their results demonstrate the effectiveness of ensemble models like Gradient Boosting in handling non-linear relationships. Chandan [17] further extends this direction by comparing Linear Regression, Decision Trees, and Passive Aggressive Regressors for reach prediction. Their findings show that engagement metrics such as likes, comments, shares, and impressions are strong predictors of reach, while also highlighting the trade-offs between model complexity, interpretability, and robustness. Kundu [18] presents a practical machine learning pipeline that predicts engagement using metadata-driven features and integrates the model into a user facing system. Their work demonstrates that even simple regression models can provide meaningful insights, though limitations such as small datasets and lack of visual or semantic analysis remain [19]. Complementary to predictive studies, Curtis focus on content credibility and psychological impact, showing that trustworthy and non-harmful content fosters sustained engagement. While not predictive in nature, this work underscores the qualitative factors that indirectly influence engagement behavior and reach. Research on emerging influencer types further expands the understanding of engagement dynamics. Silva [20] analyzed virtual influencers and demonstrate that narrative coherence, authenticity perception, and persona alignment significantly shape user interaction. Their findings suggest that engagement is influenced not only by numerical metrics but also by storytelling and audience perception.

Additionally, Yew [21] propose engagement rate algorithms that weight interactions differently, emphasizing that comments and reach provide deeper insight into influence than raw follower counts. apply statistical and machine learning techniques to study correlations between reach, engagement, and follower growth, confirming that reach is closely tied to interaction intensity rather than follower volume alone. The reviewed literature establishes that Instagram reach is shaped by a combination of engagement metrics, content characteristics, timing, and platform behavior. While prior studies contribute valuable insights through descriptive analysis, forecasting, content evaluation, and regression modelling, many focus on isolated aspects of the problem. This project builds upon these foundations by integrating engagement driven machine learning models with systematic exploratory analysis to predict reach and provide actionable insights, thereby addressing the gap between descriptive analytics and predictive, strategy-oriented Instagram reach analysis.

III. Dataset Description

The dataset used in this project consists of 104 Instagram posts collected from an organizational Instagram account and represents detailed post-level engagement analytics over a defined time period. Each record contains 12 attributes covering multiple dimensions of content performance, including identifiers and content metadata such as Post ID, Media Type, Post Type, and Timestamp. Core visibility measurement is captured through the Reach attribute, while audience engagement is reflected using Likes, Comments, Shares, Saved, and Total Interactions. In addition, textual fields such as Caption and Hashtags are included to represent content semantics and discovery mechanisms.

On average, captions contain around 22 words, indicating moderately descriptive content intended to convey context or promotional information. The dataset includes 263 unique hashtags, highlighting varied topical coverage and hashtag strategies used to improve content discoverability. The mean reach value across posts is approximately 1,948, indicating a

ISSN: 3107-6513

balanced mix of low-performing and high-performing posts, which is suitable for supervised learning tasks. Overall, the dataset captures both quantitative engagement behaviour and qualitative content characteristics, enabling comprehensive exploratory analysis, feature importance evaluation, and effective machine-learning-based prediction of Instagram reach.

TABLE I. DATASET STATISTICS

Sl.no.	Dataset Statistics	Values
1	Data Size	104
2	Number of Attributes	12
3	Features Present	Post ID, Media Type, Post Type, Timestamp, Reach, Saved, Shares, Total Interactions, Likes, Comments, Caption, Hashtags
4	Avg. Caption Length	22 words
5	Total Number of Hashtags	263
6	Avg. Reach	1948

IV. PROPOSED METHODOLOGY

The methodology adopted in this project follows a systematic machine learning pipeline, as illustrated in fig. 1, which represents the end-to-end workflow for Instagram reach analysis. The flow chart visually outlines the progression from raw data extraction to predictive modelling and insight generation. As shown in the fig. 1, the process begins with attribute extraction, where Instagram post level data is collected using unique identifiers such as Post ID. These identifiers are used to fetch essential metadata including captions, timestamps, and engagement metrics. The extracted information is initially stored in structured CSV files ensuring organized and traceable data storage. Captions and related textual attributes are harvested during this stage to support later feature engineering and text analysis. The extracted data is then consolidated into a unified dataset, which serves as the input for Exploratory Data Analysis (EDA), as indicated in the flow chart. During EDA, statistical summaries and visualization techniques are applied to understand data distribution, detect anomalies, identify outliers, and analyze relationships between reach and engagement metrics such as likes, comments, shares, and saves. This step is critical for gaining insights into user interaction patterns and for guiding feature selection.

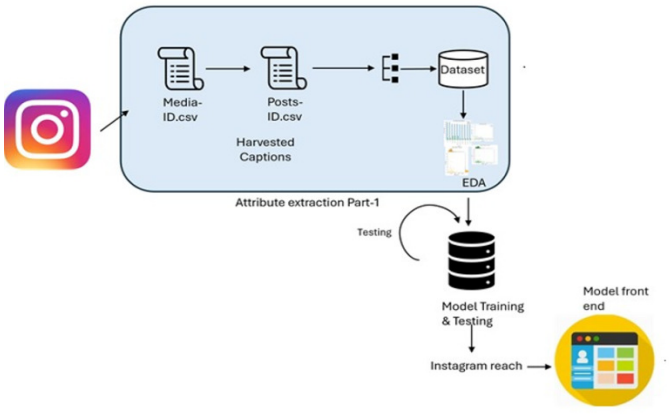

Fig 1. Proposed Flowgraph

Following the EDA, the processed dataset is passed to the model training and testing phase, shown in the central part of the Figure. Here, multiple machine learning regression models are trained using an 80:20 train-test split. The mod els learn the relationship between engagement features and Instagram reach. Testing is performed iteratively to evaluate generalization capability and to avoid overfitting. Performance is assessed using standard regression metrics such as MAE, RMSE, and R^2 score. Once the optimal model is selected, the trained system pro duces Instagram reach predictions, which are then delivered to the model front end, as depicted in the final stage of the figure. This front-end layer enables users to visualize predicted reach values and interpret insights derived from the model. The integration of analytical modelling with a user facing interface ensures that the system is not only technically sound but also practically usable for creators and marketers. The methodology depicted in the fig.1 demonstrates a complete and scalable workflow that transforms raw Insta gram engagement data into meaningful, predictive insights. By combining structured data extraction, in depth analysis, robust machine learning models, and intuitive visualization, the proposed system effectively supports data-driven decision making for optimizing Instagram content reach.

V. RESULTS

This section presents the outcomes obtained from the exploratory data analysis and machine learning experiments conducted on the Instagram engagement dataset.

A. *Data Interpretation*

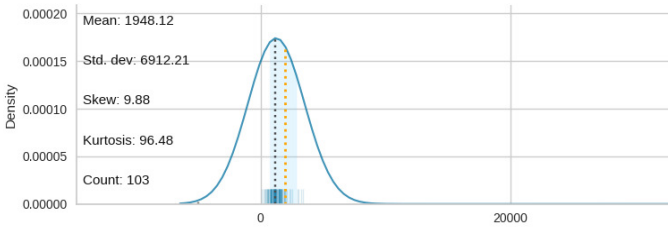

Fig 2. Distribution of the reach attribute

The distribution in fig. 2 shows that Instagram post reach varies widely across the dataset, with a small number of posts achieving exceptionally high visibility compared to the average. The strong right skewness and presence of extreme outliers indicate that viral posts heavily influence overall reach patterns rather than uniform performance across all posts.

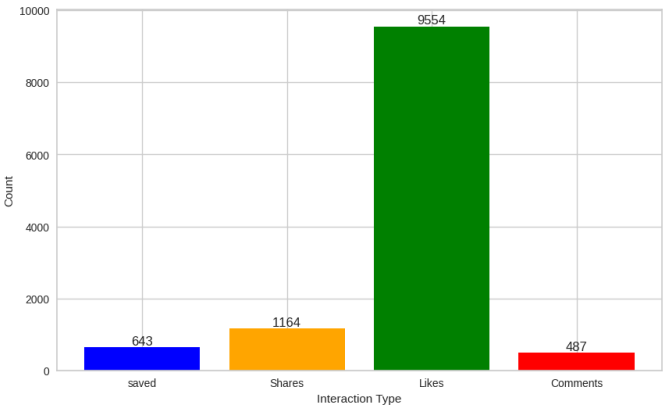

Fig 3. User Engagement Metrics

 ISSN: **3107-6513**

The fig. 3 shows that likes account for the majority of user interactions, followed by shares and saves, while comments occur far less frequently. This indicates a preference for low effort engagement and highlights likes and shares as key contributors to increased Instagram reach.

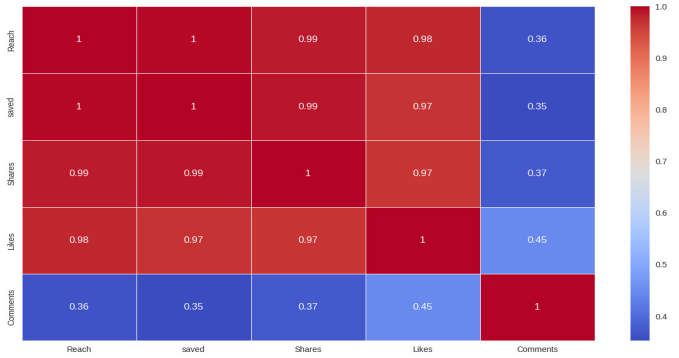

Fig 4. Correlation Analysis of Engagement Metrics

This fig. 4 demonstrates a strong positive correlation between reach and likes, shares, and saves, indicating their major role in content visibility. Comments show a weaker relationship, suggesting they contribute less to reach compared to other engagement actions.

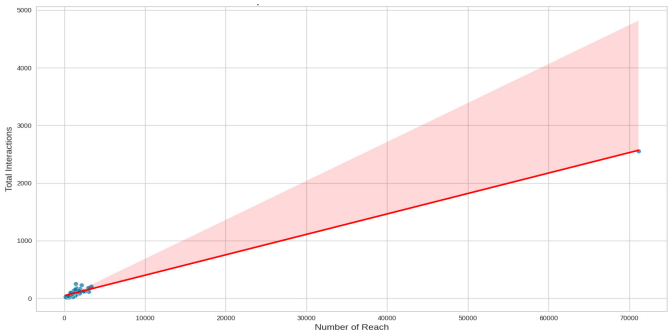

Fig 5. Relationship Between Total Interactions and Reach

The fig. 5 demonstrates a strong linear relationship between total interactions and reach, where higher exposure leads to increased engagement. High-reach outliers further confirm that engagement metrics strongly influence Instagram's content amplification.

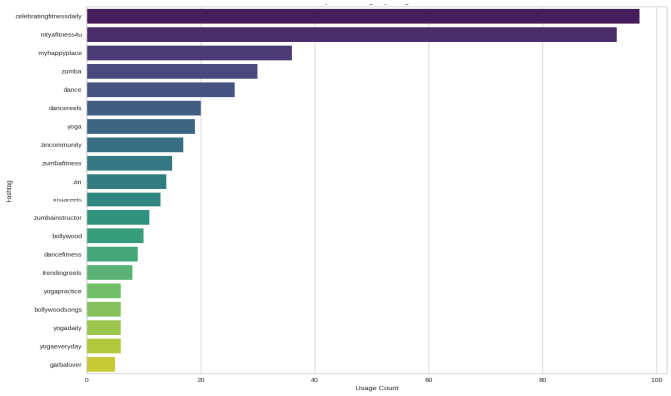

Fig 6. Top 20 Hashtags

Frequently used hashtags reflect a strong focus on fitness and activity-based themes which is depicted in fig. 6. Consistent use of niche-relevant hashtags supports targeted discoverability and improved reach.

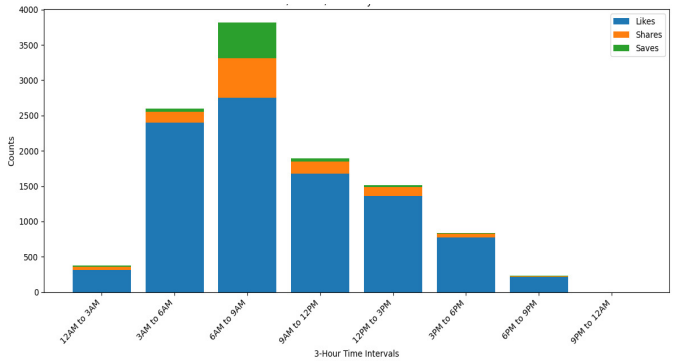

Fig 7. Engagement Distribution Across Time Intervals

As shown in fig. 7 engagement peaks between 6 AM and 9 AM, indicating this as the most effective posting window. Late-night and evening hours show significantly lower interaction levels.

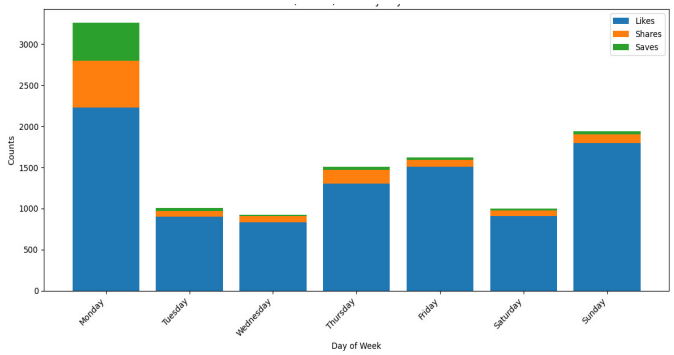

Fig 8. Engagement Distribution over Week days

Engagement is highest on Monday and Sunday, with reduced activity mid-week as per the plot shown in fig. 8. Posting at the start or end of the week may enhance reach and interaction.

B. Performance Analysis

The experimental results indicate that Instagram reach exhibits a predominantly linear relationship with engagement features such as likes, comments, saves, and shares. Linear Regression performed effectively as a baseline model, achieving an R^2 score of 0.656, which suggests that approximately two thirds of the variance in reach can be explained through direct proportional increases in user interactions. Lasso Regression produced similar results (R^2 = 0.639), reinforcing the observation that all engagement metrics contribute meaningfully to reach prediction. In contrast, tree-based ensemble models such as Random Forest and Gradient Boosting performed poorly, yielding low R^2 scores due to their inability to extrapolate beyond the training data range. This limitation proved critical in handling viral posts with unusually high reach values, resulting in large prediction errors. The TheilSen Regressor emerged as the most reliable

model, achieving the highest accuracy with an R^2 score of 0.750. Its robustness to extreme outliers, achieved by estimating trends using median based statistics rather than mean values, allowed it to focus on typical post behavior and deliver more stable predictions. Overall, the results demonstrate that robust linear models are better suited for Instagram reach prediction in datasets characterized by high variance and occasional viral spikes.

TABLE II. MODEL PERFORMANCE COMPARISION

Sl .no.	Model	R^2 Score	MAE	RMSE
1	TheilSen Regressor	0.7501	1942.55	7920.10
2	Linear Regression	0.6564	2278.81	8921.52
3	Lasso Regression	0.6395	2324.72	9138.32
4	Random Forest	0.0099	3706.45	15294.94
5	Gradient Boosting	0.0143	3780.19	15327.75

C. Key Insights

Based on the model outputs and the exploratory data analysis conducted, several important insights were identified that help explain the factors contributing to higher Instagram reach. These findings highlight behavioral patterns within the dataset and provide practical recommendations for improving content performance

- **Posting Time Impact**: Posts published between 6 AM and 9 AM consistently achieved higher reach and engagement, indicating that early posting during peak user activity increases the likelihood of rapid initial interactions and improved algorithmic visibility.
- **Hashtag Usage**: Posts using a moderate number of relevant hashtags (8–15) performed better than those with very few or excessive tags, highlighting the importance of balanced and targeted hashtag selection for improved discoverability.
- **Content Type Influence**: Reels and carousel posts demonstrated superior reach compared to single-image posts, benefiting from higher interactivity and Instagram's preference for short-form video content.
- **Engagement Correlation**: A strong positive relationship was observed between reach and engagement metrics, particularly likes, comments, and saves, with saves emerging as a key indicator of perceived content value.
- **Caption Length**: Medium-length captions (approximately 50–100 words) resulted in better engagement and reach by providing sufficient context while maintaining readability.

VI. CONCLUSION & FUTURE SCOPE

Our project analyzed the key factors influencing Instagram post reach and developed a machine learning model to predict reach based on engagement metrics and posting behavior. Through systematic data preprocessing, exploratory analysis, and model evaluation, the study identified clear patterns affecting content visibility. The results confirm that Instagram reach is strongly influenced by media type, early user engagement, posting time, and hashtag strategy rather than occurring randomly. The final predictive model demonstrated reliable performance and highlighted the importance of features such as engagement rate and content format. Overall, the project shows that machine learning can effectively support data driven content strategies and help creators improve visibility by aligning posts with observed engagement patterns.

Future Scope

- **Expanded Dataset**: Using a larger and more diverse dataset with multiple accounts, demographic details, and longer historical timelines can improve model generalization and prediction accuracy.
- **Advanced Learning Models**: Incorporating deep learning techniques such as Long Short-Term Memory (LSTM) networks for temporal analysis and Convolutional Neural Networks (CNNs) for visual content evaluation can help capture complex patterns influencing reach.
- **Caption Sentiment Analysis**: Applying Natural Language Processing (NLP) methods to analyze caption sentiment, tone, and semantic relevance can provide deeper insights into how language impacts user engagement.
- **Real-Time Prediction System**: Integrating the model into a real-time application can allow creators to estimate post reach before publishing and dynamically optimize content strategies.
- **Viral Content Detection**: Developing specialized models to detect early indicators of virality can help identify posts likely to achieve exceptionally high reach.
- **Automation via Instagram API**: Automating data ex traction, analysis, and reporting using the Instagram Graph API can enable continuous performance monitoring and scalable deployment.

REFERENCES

[1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the chal lenges and opportunities of social media," Business Horizons, vol. 53, no. 1, pp. 59–68, 2010.

[2] C. C. Aggarwal, Social Network Data Analytics. Springer, 2011.

[3] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge University Press, 2004.

[4] R. G. Curtis, I. Prichard, G. Gosse, A. Stankevicius, and C. A. Maher, "Hashtag fitspiration: credibility screening and content analysis of insta gram fitness accounts," BMC Public Health, vol. 23, no. 421, pp. 1–14, 2023.

[5] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. Gutmann, "Computa tional social science," Science, vol. 323, no. 5915, pp. 721–723, 2009.

[6] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, pp. 273–297, 1995.

[7] Meta Platforms Inc., "How instagram feed, stories, reels, and explore ranking works." https://about.instagram.com/blog, 2023. Official de scription of Instagram ranking systems.

[8] Meta Platforms Inc., "Instagram graph api documentation." https://developers.facebook.com/docs/instagram-api, 2024. Accessed: 2025-01-10.

[9] Meta Platforms Inc., "Understanding reach and impressions on insta gram." https://business.instagram.com/creator, 2023. Metrics definitions for business and creator accounts.

[10] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

ISSN: **3107-6513**

[11] P. Bellavista, L. Foschini, and N. Ghiselli, "Analysis of growth strategies in social media: The instagram use case," in 2019 IEEE, IEEE, 2019.

[12] W. H. Alwan, E. Fazl-Ersi, and A. Vahedian, "Identifying influential users on instagram through visual content analysis," IEEE Access, vol. 8, pp. 169594–169603, 2020.

[13] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5 32, 2001.

[14] M. Nemade, Y. Parulekar, A. Patel, A. Prajapati, and C. Rane, "Insta gram reach analysis," International Journal of Advanced Research in Science, Communication and Technology, vol. 5, no. 7, pp. 369–373, 2025.

[15] P. Srushti, "Instagram reach forecasting," International Research Journal of Modernization in Engineering, Technology and Science, vol. 6, no. 7, pp. 3396–3409, 2024.

[16] W. Musu, I. Samsie, A. Bastiatul Fawait, N. Lempan, Nurliah, M. M. Gabriel, and Sinar, "Trend and correlation analysis of instagram activity using data mining and statistics," in 2024 6th International Conference on Cybernetics and Intelligent System (ICORIS), pp. 1–6, IEEE, 2024.

[17] T. D. Chandan, C. S. Bhavana, J. Dilip, B. D. Shridhar, C. M. Girish war Reddy, and G. Chaitanya, "Performance analysis of machine learn ing algorithms for instagram post reach analysis," in 2023 International Conference on Computational Intelligence for Information, Security and Communication Applications (CIISCA), pp. 360–365, IEEE, 2023.

[18] R. Kundu, S. Ghosh, S. Shreyansh, Y. Yadav, and B. Rao, "Instagram reach analysis and prediction," International Journal of Innovative Research in Technology, vol. 9, no. 12, pp. 952–956, 2023.

[19] D. Chaithanya, P. Hemashree, B. D. Shridhar, A. Esha, and J. S. ImpuD, "Text emotion detection using machine learning algorithms," in 2023 International Conference on Computational Intelligence for Informa tion, Security and Communication Applications (CIISCA), pp. 304–306, IEEE, 2023.

[20] M. J. de Brito Silva, L. d. O. R. Delfino, K. A. Cerqueira, and P. d. O. Campos, "Avatar marketing: a study on the engagement and authenticity of virtual influencers on instagram," Social Network Analysis and Mining, vol. 12, no. 130, pp. 1–19, 2022.

[21] R. L. H. Yew, V. K. Sevamalai, S. B. Suhaidi, and P. Seewoochurn, "Social network influencers' engagement rate algorithm using instagram data," in 2018 IEEE Conference on Big Data and Analytics, pp. 1–8, IEEE, 2018