IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14816

Predictive Analytics for Social Media Engagement

Aishwarya M K[1], Dr. Hemanth Kumar[2], Dr. Ashwini J P[3]

Student, Department of MCA, Jawaharlal Nehru New College of Engineering, Shivamogga, India[1]

Associate Professor, Department of MCA, Jawaharlal Nehru New College of Engineering, Shivamogga, India[2]

Associate Professor, Department of AI&ML, Jawaharlal Nehru New College of Engineering, Shivamogga, India[3]

Abstract: Initiating and involving oneself in social media has become involved in all people's lives today. It allows people to connect with people in different places and share content, information, experience, ideas, and more. Since so many people take part in these activities every day, businesses and organizations are getting into the market as well by marketing, advertising, promoting their brands, and getting more clients. It is also possible to analyse user activity with post engagement. This work used datasets of different features from the past to understand engagement of post and applied Machine Learning (ML) methods to analyse and interpret user activity and measure the amount of engagement of users. The datasets included caption and post time, media type, post length, CTR (click-through rate), ad-interaction time, and hashtag. The datasets also used Machine Learning (ML) algorithms to predict how many interactions a post may get. The results are displayed in graphs that show cluster of users and the engagement activity of each post.

Keywords: Machine Learning, Social Media, Engagement Metrics, Predictive Analytics.

I. INTRODUCTION

Social media serves as a common part of life and as a need for businesses, marketing, and communication. Platforms like YouTube, Twitter, Instagram, Facebook, and others play a far greater role than just personal use as many of them use social media as a tool in advertising and influencing marketing. Social media's increasing popularity has produced an ever-growing amount of data on users. In this work, we choose Instagram platform for analysing predicting the post engagement by using historical data of Instagram. The various features are included in dataset such as caption length, caption, media type, likes, shares, comments and more to analyse and predict post engagement in social media. It also includes metrics like click-through-rate, ad interaction time adds importance to analyse the behaviour of user. we used ML techniques to analyse and predict post engagement in social media such as Random Forest and Linear Regression used to forecast engagement of post and feature importance. Random forest works for Classification and Regression tasks. K-Means is an unsupervised clustering helps to group the users based on the user behaviour for targeting the audience. The outcomes are visualised using graphs using matplotlib and also showed the accuracy comparison graph of random forest and linear regression. This work is helpful for businesses, digital marketing to understand their customers and improve their strategy to reach a greater number of audiences.

II. LITERATURE SURVEY

Andrea Ochsner et al. developed a model for analysing social media data through an extensive literature and used twitter data [1]. They differentiated between various terms used in social media analysis and focused on developing, adapting, extending informatics tools, methods to track, collect. They analyse the large amounts of structured, semi-structured and unstructured data and information of social media. Authors in [2] developed a model that uses twitter data and removes duplicates and develop predictive applications of user generated content on twitter. There is no clear methodology mentioned and uses twitter API constraints that allows to extract only limited number of tweets. Work by Edyta Golab aims to highlight the significance of artificial intelligence and its tools in enhancing customer engagement within social media [3]. He analyses the customer engagement in social media. Rakesh Kundu et al. designs a model to analyse Instagram data and make predictions based on the machine learning techniques [4]. They work is done based on the steps of collection of data, preprocessing, train the selected model, evaluate trained model, and predict result. Authors in [5] have proposed a model that utilised social media data by integrating feature selection, model comparison, cross-validation and data visualisation. The activities in social media have been considered as a predictor of the success of movies in box office. Work by Xingting Ju proposed a model of studying the impact of COVID-19 pandemic on the behaviour of social engagement, who noted the difficulties it posed to marketing approaches and the necessity to enhance online communication with clients [6]. It analysed Twitter pages of 23 well-known American catering customers and observed

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14816

the pre-pandemic era and pandemic era. It emphasizes food and lifestyle, marketing protocols and online food ordering. Ioannis C Drivas et al examined unstructured data from social media provided by the users [7]. They identify and analyse the currently used metrics in social media analytics, categorizing them according to their purpose. Their approach follows a process of determining information needs, collecting data, and organizing, analysing, interpreting, and acting on the data as information, to best provide meaningful insights into user engagement. Authors in [8] will be using multiple social media data sources like wikis, blogs, chats, news feeds, and multimedia. They used web API to process twitter data, analyse data through sentiment analysis and interpret data. Kayli Blackburn et al. developed a model aims at explaining engagement among viewers on social media [9]. Among the demographic and personal characteristics, considered by the authors, are: education, gender, place of residence, age, and personal interests. The sentiment of users can be analysed using AI and ML with regard to some of these features, therefore, the model can give an insight on the many aspects related to the engagement of the audience in social media. Authors in [10] designed a model based on the previous social media information, which is under the scientific approach to data collecting, analysing and evaluating with the help of a great number of tools and technologies and they observe that the particular concern is related to data privacy alongside the technical problems. Abdulfattah Ba Alawi et al. introduced a model that uses Twitter data to derive meaningful information that can be used in making informed decisions [11]. These are characterized as a combination of ML techniques that create a hybrid model with increased predictive capability. The steps that were described in this paper can be broken down into several steps that might be related to the identification of data sources, sentiment analysis, the model's usage, the framework creation, as well as its evaluation. Work by Sumit Jain examine customer activities, preference and opinion in different social media sites [12]. This paper explains how a strategy of this type can enable brands to track critical engagement indicators, such as likes, comments, and shares, which directly influence customer decision-making activities. The model allows analysis of sentiment, which helps to gain better insight into customer attitudes and contributes to customer behaviour, in general. Authors in [13] designed a model to enhance the operations of US fashion brand types and analysed modifying dynamics in the changing digital space, social media analytics and customer engagement. The process in this paper consisted of collection of data and pre-process data, feature engineering, model validation, evaluation, and visualization. The study results showed the predictive models were successful at estimating and optimizing brand performance in the fashion industry. Prakrit Saikia et al. developed a model for evaluating the social media posts made by higher education institutions [14]. This research aimed to explore various characteristics of posts and the engagement levels inferred from a sample of 29,814 observed from either Facebook or Instagram. The approach taken in this study was based on sampling, systematic data collection, analytical equations and categorization of content that had been collected. Authors in [15] proposed a model that gathers features that is useful for engagement of user in Instagram. The features in the dataset are age of creator, count of followers, gender, post day and time, words, emojis, hashtag. They use ML technique like LASSO regression for 13000 posts of German speaking influencers. They identify many factors such as context factor, content factor, posting on evening of Friday, use emoji's, people's photos and any text.

III. METHODOLOGY

This work involves Machine Learning and Natural Language Processing techniques as follows:

A. Machine Learning (ML)

Machine Learning is a way that the computers learn from the data and make the prediction on their own not being definitely programmed. As like how humans learns from experience as like Machine Learning enhance performance when we give more information to work with on. It has three types, namely: Supervised, Unsupervised and Reinforcement Learning. We discussed Random Forest Classifier, Random Forest Regressor, Linear Regression and K-Means Clustering.

Algorithm	Purpose	Role in work
Random Forest Classifier	We use this for forecasting continuous values like engagement of audience and works better with regression tasks.	In this work, this used to know the engagement of posts uploaded in future.
Random Forest Regressor	We use this for divide the audience like high or low engaged.	In this work, this used to classify as high or low engaged audience.
Linear Regression	We use this to identify the relationship of training and testing values by fitting.	In this work, this used to forecast the metrics of engagement of posts.
K-Means Clustering	We use this to apply clustering on users and divide them into different groups with help of clustering technique.	In this work, this used to segregate users based on their interest and group into four groups.

Table.1 Machine Learning Methods

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14816

B. Natural Language Processing (NLP)

Natural Language Processing is used to identify and understand human language. It read text by human, proceed to computer system and convert information to human text. In this paper, Natural Processing Language is employed to analyse the text in contents of social media. It aids in transforming the unstructured captions into a structured features like hashtags, length of captions, the emoji practice, the question marks practice. It provides text-based features to Machine Learning models in order to predict the post engagement.

Technique	Purpose	Role in work
Valence Aware Dictionary and sEntiment Reasoner (VADER)	It is the method of Natural Language Processing that is used to analyse the sentiment by assigning the polarity values between -1 and +1.	In this work, this method is employed to determine score of sentiment of every caption in post. The analysis of sentiment score feature helps to check the emotion of user to a post. It recognizes the feeling of user like positive, neutral or negative.
Latent Dirichlet Allocation (LDA)	It is a method applied in Natural Language Processing that assists in carrying out topic modelling on social media post captions.	In this work, this method applied to extract the topics covered in the caption so that it classified into different groups including food, travel, fashion, quotes and fitness.

Table.2 Natural Language Processing Techniques

C. Block Diagram

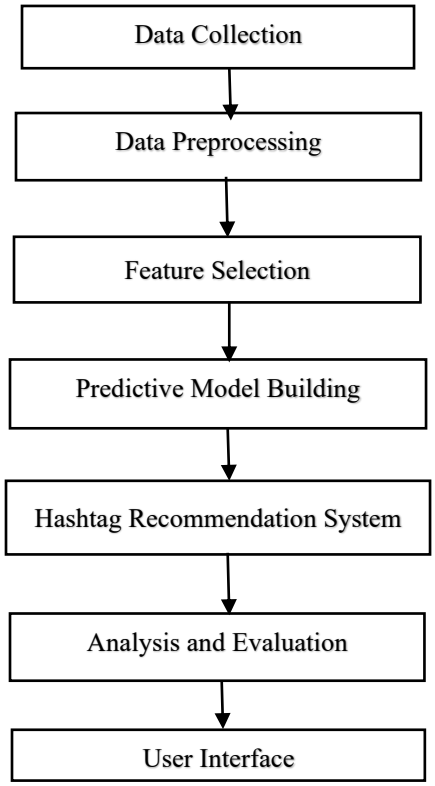

Fig.1 Block Diagram

The data that is used in this work in step 1 contains information about social media posting and other parameters. It contains the caption material, media type(image, video, reel, carousel), likes, shares, comments, saves, click-through-rate, ad-interaction-time and other associated users and post data. The preprocessing in step 2 entails data cleaning, the transformation of categorical values into the numerical values, null values or dropping null values, missing values, and feature extraction to achieve the results. The key attributes are chosen such as captions, sentiment score, hashtags, presence of questions, presence of emojis, media type and time-based features to recommend the time to post the content, click-through-rate, ad-interaction-time to know how many seconds the users spent on the ad in step 3. Step 4 calculates the engagement metrics consist likes, comments, shares, saves and will project the engagement post. The Random forest classifier is employed to label the posts as either high or low engagement, whereas the Random forest regressor and linear

DOI: 10.17148/IJARCCE.2025.14816

regressor are utilised to estimate the engagement indicators. In user grouping, K-Means clustering algorithm can be useful in grouping the users into meaningful classes. The sentiment of captions is analysed by using Valence Aware Dictionary and sEntiment Reasoner (VADER), and the Latent Dirichlet Allocation (LDA) topic modelling is used to identify the underlying themes in the captions. The system in step 5 also features a rule-based Hashtag Recommendation which will make the accurate recommendations of hashtags based on type of media and the caption typed by user. These suggestions aid in the spread of the posts and enhancing the engagement with users. In step 6, the performance of model is assessed by accuracy and other measures and the results slowed in visualisations. Upon that the impact of such features as whether a caption included questions or emojis is also taken into consideration. Lastly, in step 7, an interactive interface is created with Streamlit in which a user can enter a caption and media type in order to generate hashtag suggestions. The results, as well as the engagement statistics are displayed in form of graphs generated with matplotlib, which makes the system convenient and visually comprehensible.

IV. NUMERICAL ANALYSIS

The evaluation of model evaluated using accuracy, precision, recall and f1-score for Random Forest Classifier as shown below:

A. Accuracy: It finds out the percentage of correctly predicted values to the total values. It tells correctness of engagement classification overall.
Equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

In equation (1),
TP = True Positives that is correctly predicted positive values.
TN = True Negative that is correctly predicted negative values.
FP = False Positive that is incorrectly predicted as positive.
FN = False Negative that is incorrectly predicted as negative.

B. Precision: The part of number of positive values which are predicted to the number of positive values. For example, it shows how reliable the engagement predictions are.
Equation:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

In equation (2),
TP = True Positives.
FP = False Positives.

C. Recall: It is an estimated positive value that are accurately estimated out of the positive values. For example, it shows how well the model predicts actual engagement.
Equation:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

In equation (3),
TP = True Positives.
FN = False Negatives.

D. F1-score: It is a mean precision and recall. It balances both and gives single performance indicator.
Equation:

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

In equation (4),
Precision = values of equation (2).
Recall = values of equation (3).

	precision	recall	F1-score
0	0.96	0.86	0.91
1	0.92	0.98	0.95
Accuracy			0.93
macro avg	0.94	0.92	0.93
Weighted avg	0.94	0.93	0.93

Table.3 Performance Evaluation of Model

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14816

V. RESULTS AND DISCUSSION

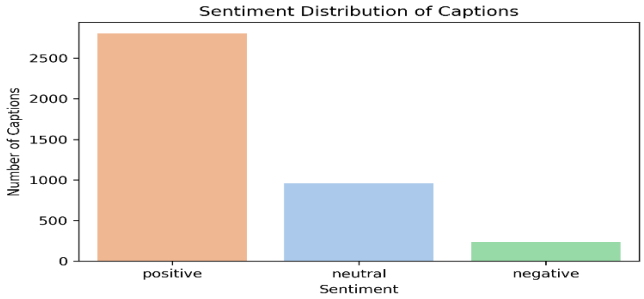

Fig.2 Sentiment Distribution of captions

Fig.2 shows the Sentiment Distribution expressed in captions by the audience. There are three bars represents positive, neutral and negative sentiments on captions. Each bar showed in different colours like blue, orange and green. As shown in graph, the positive sentiment is more than other sentiments.

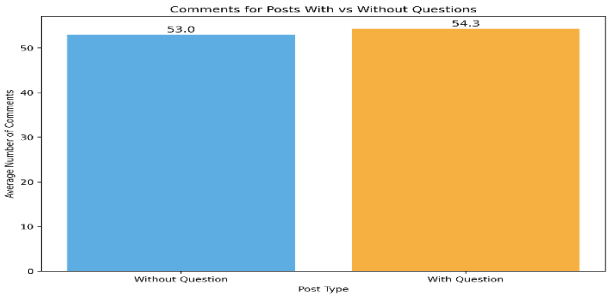

Fig.3 Comments for posts with vs without questions

Fig.3 shows the analysis of Comments for posts with or without the questions type in captions. Caption with questions shows in orange colour and caption without questions shows in blue colour. While the post with questions gets more comments as compared to without questions.

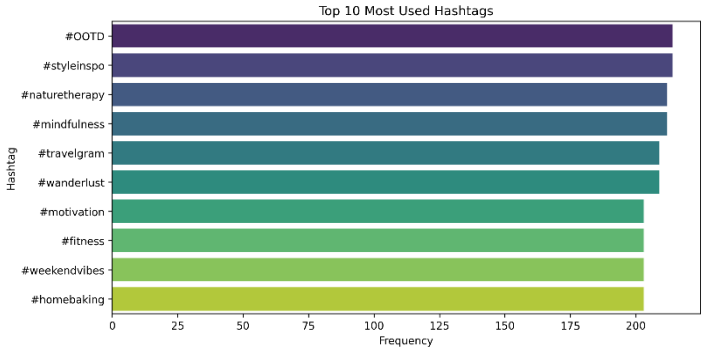

Fig.4 Top 10 most hashtags

Fig.4 shows the most used hashtags. The frequently used hashtags are #OOTD, #styleinspo, #naturetherapy, #mindfulness, #travelgram, #motivation, #fitness, #weekendvibes and #homebaking. These are the hashtags widely used the users.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025

DOI: 10.17148/IJARCCE.2025.14816

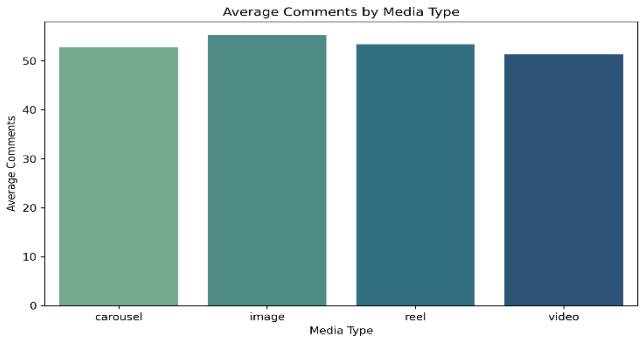

Fig.5 Average Comments by Media Type

Fig.5 shows the Average Comments by Media Type such as carousel, image, reel and video. The x-axis shows the media type and the y-axis shows the average comments.

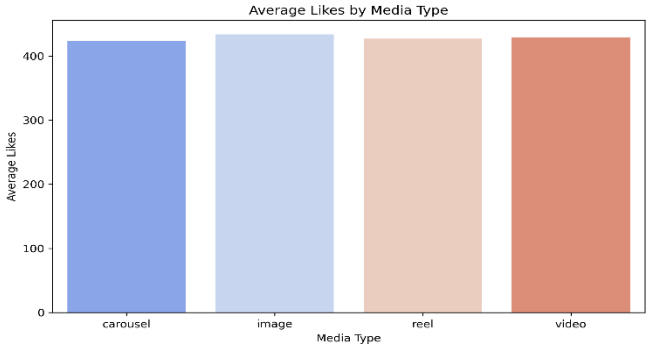

Fig.6 Average likes by Media Type

Fig.6 shows the Average Likes by Media Type. The type of media are carousel, image, reel and video. The x-axis represents the media type and the y-axis represents the average Likes.

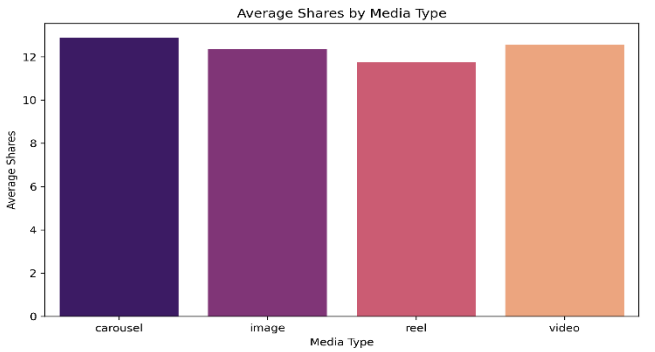

Fig.7 Average Shares by Media Type

Fig.7 shows the Average Shares by Media Type. The type of media are carousel, image, reel and video. The x-axis represents the media type and the y-axis represents the average Shares.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14816

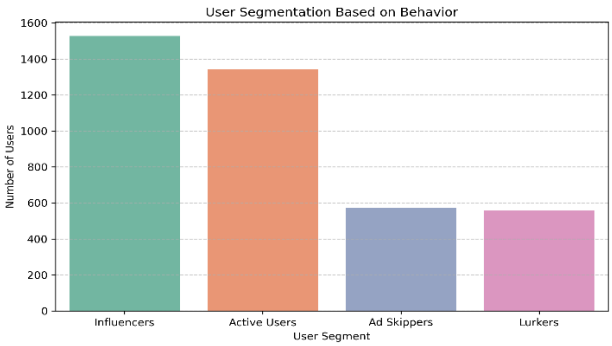

Fig.8 User Segmentation

Fig.8 shows the user segmentation into different type of users based on behaviour. The dissimilar users are clustered into influencers, active users, ad skippers and lurkers. The largest segment is influencers followed by active users. Each user segment shows in different colours.

VI. CONCLUSION

Social media apps are playing very significant part in digital marketing to reach the wider number of people. This work shows how predictive analytics can increase demand of content of social media and predict engagement metrics. By mixing Machine Learning methods such as random forest algorithm and k-Means clustering and also VADER tool is aimed at sentiment analysis. It can predict engagement metrics based on the input given by the user and also recommend hashtag suggestions based on media type. This work also include a user interface built using Streamlit to display all the results and graphs to the users and also can view the hashtag suggestions. Overall, this work combines the Machine Learning models, study sentiment score, Topic Modelling based on captions, Hashtag Recommendation, Engagement Predictions and User Interface. This work is helpful for content creators, business and digital marketing people to improve their business and reach larger number of audiences.

REFERENCES

[1]. C. Zachold, O. Samuel, A. Ochsner, and S. Werthmuller, “Analytics of social media data-state of characteristics and application,” Journal of Business Research, Vol. 144, pp. 1064-1076, Feb. 2022.
[2]. E. Cano Marin, M. Mora Cantallops, and S. Sanchez Alonso, “Twitter as a Predictive system: A systematic literature review,” Journal of Business Research, Vol. 157, Dec. 2022.
[3]. E. Golab Andrzejak, “Enhancing Customer Engagement in Social Media with AI – A Higher Education Case Study,” Procedia Computer Science, Vol. 207, pp. 3022-3031, 2022.
[4]. 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, May. 2023.
[5]. S. L. Mary and S. B. Deepthi, “Integrating predictive analytics and social media,” *Journal of Emerging Technologies and Innovative Research*, Vol. 5, No. 9, pp. 349-358, Sep. 2018.
[6]. X. Ju, “A Social media competitive intelligence framework for brand topic identification and customer engagement prediction,” *PLOS ONE*, Vol. 19, No. 11, Nov. 2024.
[7]. I. C. Drivas, D. Kouis, D. Kyriaki Manessi, and F. Giannakopoulou, “Social Media Analytics and Metrics for Improving Users Engagement,” *Knowledge*, Vol. 2, No. 2, pp. 225-242, May. 2022.
[8]. B. Batrinca and P. C. Treleaven, “Social media analytics: a survey of techniques, tools and platforms,” *AI & Society*, Vol. 30, No. 1, pp. 89-116, 2015.
[9]. K. Blackburn, K. Boris, E. Frankum, and G. Zornes, “Social Media Data Analytics-Using Big Data for Big Consumer Reach,” *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Vol. 10, May. 2022.
[10]. B. Patra and A. Mahalwar, “Social media analytics: techniques and applications,” *International Journal of Mechanical Engineering*, Vol. 6, No. 1, pp. 1696-1709, Nov.-Dec. 2021.
[11]. A. B. Alawi and F. Bozkurt, “A hybrid machine learning model for sentiment analysis and satisfaction assessment with Turkish universities using Twitter data,” *Decision Analytics Journal*, Vol. 11, Apr. 2024
[12]. S. Jain, “Optimizing Customer Engagement with Social Media Analytics,” *International Journal of Advanced Research and Multidisciplinary Trends (IJARMT)*, Vol. 1, No. 1, pp. 41-45, Jul.-Sep. 2024.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 8, August 2025
DOI: 10.17148/IJARCCE.2025.14816

[13]. M. R. H. Ontor, A. Iqbal, E. Ahmed, Tanvirahmedshuvo, and A. Rahman, "Leveraging digital transformation and social media analytics for optimizing US fashion brands' performance: A machine learning approach," *International Journal of Computer Science & Information System*, Vol. 9, No. 11, pp. 45-56, Nov. 2024.
[14]. P. Saikia and H. Barman, "A Systematic Analysis of Higher Educational Content Over Social Media for Engagement Optimization," *Review of Marketing Science*, Vol. 21, No.1, pp. 77-110, 2023.
[15]. R. Jaakonmaki, O. Muller, and J. vom Brocke, "The impact of Content, Context, and Creator on User Engagement in Social Media Marketing," *Proceedings of the 50th Hawaii International Conference on System Sciences*, pp. 1152-1161, 2017.

BIOGRAPHY

Aishwarya M K is a student of the Master of Computer Applications (MCA) at Jawaharlal Nehru New College of Engineering (JNNCE), Shivamogga. I have recently completed my research work titled "Predictive Analytics for Social Media Engagement", which focused on Machine Learning and Natural Language Processing techniques to analyze and predict engagement metrics of social media post.

Dr. Hemanth Kumar is an Associate Professor in the Department of Master of Computer Applications (MCA), Jawaharlal Nehru New College of Engineering (JNNCE), Shivamogga. He has published articles in international journals and has presented papers in international conferences. His area of interest is Wireless Sensor Networks, IoT and Machine Learning.

Dr. Ashwini J P is working as Associate professor in the department of AIML, JNN College of Engineering. She has 20 years of teaching and 14 years of Research experience. She has pursued her Ph.D. in the area of Cloud and High-Performance Computing. Her area of interest includes Cloud and Edge Computing and AI&ML.