International Conference on Advances in Technology and Computing (ICATC 2023)

Faculty of Computing and Technology, University of Kelaniya, Sri Lanka
15th December 2023

Predictive Analysis on Social Media Content to Become Viral

Singhe Silva
Faculty of Computing and Technology
University of Kelaniya
Sri Lanka
singhesilvaaaa@gmail.com

Dr. Rasika Rajapaksha
Faculty of Computing and Technology
University of Kelaniya
Sri Lanka
rasikar@kln.ac.lk

***Abstract*—In the continuously changing world of social media, Instagram has taken a prominent place by becoming one of the most popular social media platforms. Instagram has not only the biggest organic reach but also the highest organic engagement rate. Above all, understanding and predicting what makes posts go viral is an uneasy yet significant challenge. This study focuses on Instagram and presents a fresh approach to discovering the key factors contributing to post virality, specifically on image posts. In this study, a metric named 'Virality Rate' is defined to predict the likelihood of going viral. It is calculated by dividing the sum of number of likes and comments by the number of followers. There were studies on Instagram post popularity prediction based on various features and datasets. But with a focus on public image-based posts from influencers worldwide, this research delves into sentiment analysis, image processing for technical features and content, hashtag assessment, user history and user features to forecast the potential virality of a post. This research trained and compared several regression models to predict the Viral Rate and employed Faster R-CNN and OpenCV to detect objects and help extract essential technical details. Through rigorous model training and evaluation, our results highlight the Random Forest Regression model as the most effective predictor. It boasts an impressive Mean Absolute Percentage Error (MAPE) of 0.15, which implies an accuracy of 85% and a notable R-squared (R2) value of 0.924 which is significant compared to previous studies. It was found that the User History Features, sentiment score, technical features and posting time have a high impact on Virality Rate. In conclusion, this research aims to advance social media analytics by offering actionable insights for content creators, influencers, marketers and regular users.**

Keywords—Instagram, social media, viral content, predictive analysis, Virality Rate

I. Introduction

Social media has revolutionized how people communicate, share information, and engage with each other through the internet. It confines a vast range of platforms that stimulate creating, sharing, and exchanging content, delivering individuals and organizations the ability to connect and collaborate, not only nationally but also internationally. It has shaped into a vital part of modern society, creating trends, influencing public address, and nurturing connections across geographic and cultural boundaries. While there is a wide range of social media platforms, Instagram offers distinctive advantages that make it an exceptional candidate for studying the virality of social media content. Its impressive organic reach is one of its standout features, with a reported engagement rate of 9.4%. This high engagement rate is connected with an organic engagement rate of 1.16%, further emphasizing the effectiveness of the platform in capturing users' attention and encouraging interaction. Moreover, the user base of Instagram displays a distinctive sense of engaging with brands, as evidenced by the fact that 36% of consumers choose Instagram to follow companies, demonstrating a higher intention of purchasing compared to other platforms. Additionally, the platform's vast usage of sharing visual content, with its user-friendly interface, contributes to its captivation as a research focus. Furthermore, Instagram's noticeable role in shaping modern social media culture makes it a fascinating subject of analysis.

Given these reasons, Instagram delivers a rich and diverse dataset to explore the sophistication of predictive analysis for content virality. Delving into the underlying mechanisms contributing to the virality of Instagram posts and developing predictive models that can shed light on the factors driving engagement is the prime aim of this research. By understanding the dynamics of content virality on Instagram, this research seeks to empower marketers, content creators, and social media influencers with insights into effective techniques for improving engagement and creating impactful content.

The challenge sought to address is the unpredictability of Instagram content virality. Despite the widespread use of Instagram and the ample desire to create popular posts, there is an absence of a clear understanding of the factors that direct to high engagement, such as likes and comments. This unpredictability hinders the ability of content creators to consistently produce impressive posts and hinders businesses' efforts to market their products or services. Hence, the prime aim of this research is to analyze and predict the virality of posts on Instagram by identifying the crucial factors contributing to high virality rates.

To address the above-mentioned problem, this research involves using data analysis and ML regression models to predict and enhance the popularity of Instagram posts. A dataset was found, and the other required attributes, like post details and features, were extracted or scraped using Instaloader [1]. By analyzing these data, patterns were identified that contribute to high virality. Then ML models that use these patterns were built to predict how popular a post would be. Additionally, OpenCV and Faster R-CNN were used for Image Processing to identify the visual aspects that attract attention. To predict virality, a metric named ‘Viral Rate’ is defined. It is calculated by dividing the sum of likes on the post and comments on the post by the number of followers. Since the number of likes and comments depends massively on the number of followers, to do a prediction fair to all the users, the sum is divided by the number of followers.

II. Related Work

Massimiliano Viola and Luca Bruneli's paper, ‘Instagram Images and Videos Popularity Prediction: A Deep Learning-Based Approach’ [2] uses machine learning, specifically CNN, to predict Instagram post popularity. They use visual content from a specific profile and propose interpretability traits. However, challenges like over-parameterization, low recall, limited dataset size, lack of data augmentation, and noise affect the model's reliability. Also, the prediction precision has been limited due to classifying popularity into two or three categories.

‘How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention’ [3] paper proposes a dual-attention model for predicting Instagram post popularity for a specific user, incorporating image, caption, and user environment. It analyzes the correlation between words, emojis, and post popularity statistically. The model has superior accuracy and F-measure results, but its assumption that only image-caption pairs is available for prediction may limit its applicability. The model's performance may also vary depending on the user and their posting habits, as it is based on a specific user.

Kristo et al.'s study, ‘Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features’ [4] uses a global dataset to predict Instagram post engagement rates, finding that Support Vector Regression (SVR) has a prediction accuracy of up to 73.1%. The study highlights the impact of image quality, posting time, and type of image on ER, with user history features and manual image assessment values as top predictors. However, the study acknowledges limitations in addressing data variability and potential biases in image assessment, hashtags, and user history features.

In addition to the content discussed above, there was research on popularity prediction on Instagram considering the features such as image content [5], image aesthetics, hashtags [6], posting time and metadata separately or as a whole. However, no studies were found that focused on all the features, including sentiment analysis, hashtag assessment, image technical features, image content, posted time and day, user features and user history features in one paper. Moreover, most existing studies have a small data variance since they have used a local dataset or a considerably small dataset. This research uses all the above-mentioned features and also a global dataset of 15,000+ data points.

III. Methodology

There were four phases in this research namely, Data Collection, Data Filtration, Feature Extraction and Virality Prediction as shown in the Fig 1.

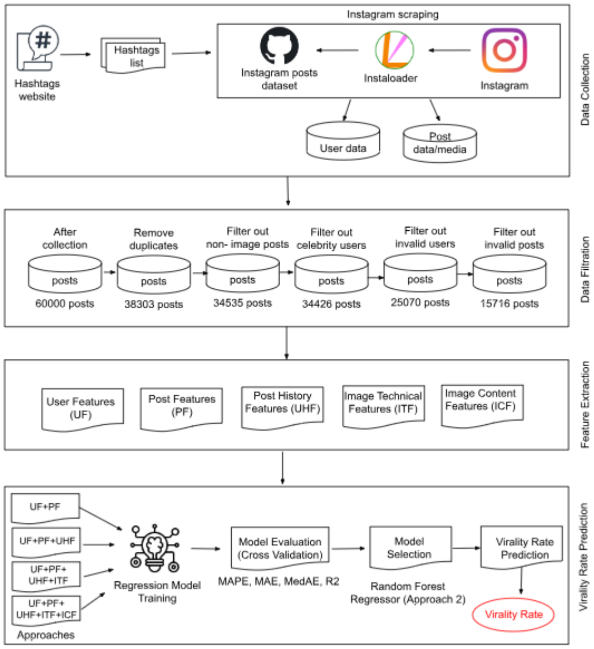

Fig. 1. *Research Methodology*

A. Data Collection

The research needed a dataset of Instagram influencers. After spending a sufficient amount of time searching a dataset on Kaggle and GitHub, two datasets were found on a GitHub repository owned by Giovanni Alcantara. The attached article written on Medium by Corentin Dugué gives a clear idea about how the dataset was created and what it was created for, which satisfies the requirement of the research [7]. Both of the datasets contain 30,000 data points each.

B. Data Filtration

Then the two datasets were concatenated vertically, which added up to 60,000 data points. Then the duplicate data points were removed from the 60,000 data points based on the ‘Post ID’ of each post. It reduced the number of posts to 38,303. The number of unique users was found to be 1,827. Since in this research, only images are considered but not videos, the data points with videos were removed. This reduced the number of posts to 34,535. Then the users with more than 1 million followers and posts with more than 200,000 likes were removed since they cannot be considered just influencers but celebrities or stars. This would also reduce the error of our model. This reduced the number of posts to 34,426.

The next challenge was about invalid posts. Since this dataset was created from the posts posted in 2016 and 2017, there is a tendency for some of the post IDs in the dataset to be invalid. If the post ID is valid, it would show the corresponding post, but if it is not valid, it would show an error message, ‘Sorry, this page isn't available. The link you followed may be broken, or the page may have been removed. Go back to Instagram.’ Firstly, the number of posts of each unique user was scraped using Instaloader, and only the valid usernames and their respective number of posts were saved in a CSV file. It was found that only 1,447 out of 1,827 were valid. Hence the data points with invalid usernames were removed from the dataset, which reduced the dataset to 25,070 data points. Then Instaloader was further used to identify the invalid posts and filter out only the valid posts based on the post IDs. This reduced the dataset to 15,716 data points.

C. Data Extraction

The list of feature categories and the extracted features are as follows:

1) User Features (UF): These features were already found in the dataset.

- num_followers: Number of followers [11,820-804,699]
- num_followings: Number of followings [0-7,500]

2) Post Features (PF): Uploaded day of the week and the hour were extracted from the 'date' column. The number of hashtags and the number of mentions were extracted from the 'caption' column. To calculate the hashtag score for each post, the most popular 10,000 hashtags were scraped from top-hashtags.com website [8]. Then a score was given to each hashtag, where the most popular hashtag was given a score of 10,000 and the least popular hashtag was given a score of 1. Then the total hashtag score was calculated by summing up the score of each hashtag. Then to calculate the sentiment score of the captions including the emojis of each post, NLTK library [9] along with VADER [10] model were used.

- day_of_week: Uploaded day of the week [1(Monday) - 7(Sunday)]
- hour: Uploaded hour of the day [1 (00:00-00:59) - 24 (23:00-23:59)]
- num_hashtags: Number of hashtags in a caption [0-30]
- num_mentions: Number of mentions in a caption [0-20]
- hashtag_score: Sum of the scores given to each hashtag [0-384,863]
- sentiment_score: Score given by sentiment analysis for the caption [0.0109-1.9913]

3) User History Features (UHF): To calculate UHF, the number of likes and the number of comments on the posts posted by the same user within the previous 30 days of the posted day of each post were captured. By using the captured number of likes, the number of comments and the already existing 'num_followers' column, the following metrics were calculated.

- mean_num_likes: Mean number of likes of the posts [17.342 – 110,294.0]
- mean_num_comments: Mean number of comments of the posts [0.0 – 2,384.267]
- av_erl: Average of (number of likes of each post/ number of followers) of the posts [0.000355 - 0.704488]
- av_erc: Average of (number of comments of each post/ number of followers) of the posts [0.0 - 0.01987]
- st_erl: Standard Deviation of (number of likes of each post/ number of followers) of the posts [0.000004 - 0.2823]
- st_erc: Standard Deviation of (number of comments of each post/ number of followers) of the posts [0.0 - 0.039065]

4) Image Technical Features (ITF): To extract the ITF, all the images corresponding to each post were downloaded using the image URLs and the 'requests' module. To extract the width, height and the size of the images, Python Imaging Library (PIL) was used. And to calculate the colorfulness, noise, sharpness, clarity and sharpness, the 'OpenCV' and 'numpy' libraries were utilized.

- image_width: Width of the image [320 – 1,279]
- image_height: Height of the image [167 – 1,600]
- image_size: The size of the image in kilobytes [6.189 – 1,368.34]
- colorfulness: Colorfulness of the image [0.0 – 3,865.077]
- noise: Noise of the image [1.775 - 146.054]
- sharpness: Sharpness of the image [1.507 – 42,285.124]
- clarity: Clarity of the image [38.358 – 230,706.601]

5) Image Content Features (ICF): Posts with images produce higher engagement than posts with only text [11]. The majority of people prefer seeing the faces of people, landscapes, animals and talent rather than seeing a blank picture or a picture with only text [12]. Hence the content of the images were considered in this category of features. For this, 'torch', 'torchvision' libraries and the pretrained 'Faster R-CNN' model were used. Faster R-CNN was used for object detection, and to label the identified objects, COCO class names were defined. Then the detected objects were categorized into seven categories as follows.

- Vehicles: car, motorcycle, aeroplane, bus, train, truck, boat, bicycle
- Animals: bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, mouse
- Food: banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, doughnut, cake, spoon, bottle, plate, wine glass, cup, fork, knife, bowl, dining table
- Sports: Frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket
- Clothing: hat, backpack, umbrella, shoe, eyeglasses, handbag, tie, suitcase, watch
- Indoor objects: chair, couch, bed, mirror, window, desk, toilet, door, TV, laptop, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, blender, book, vase, scissors, teddy bear, hair drier, toothbrush, hairbrush
- Outdoor objects: traffic light, fire hydrant, street sign, stop sign, parking meter, bench, potted plant

Seven new columns were created with the above categories, and each identified object was given a score of '1'. The total of the scores in each category was stored in each column. For example, if an image has a car, two motorcycles, and four birds, the 'vehicle_pic' column stores a score of '3', and the 'animal_pic' column stores a score of four. Hence the ICF are as follows.

- person_count: Number of people in the image [0 - 54]
- vehicle_pic: Number of vehicles related objects in the image [0 - 33]
- animal_pic: Number of animals related objects in the image [0 - 34]
- food_pic: Number of food related objects in the image [0-32]
- sport_pic: Number of sports related objects in the image [0 - 8]
- clothing_pic: Number of clothes related objects in the image [0 - 13]
- indoor_objects_pic: Number of indoor related objects in the image [0 - 49]
- outdoor_objects_pic: Number of outdoor related objects in the image [0-27]

D. Virality Prediction

To predict the Virality Rate, four main approaches were conducted by training a couple of regression models. In Approach 01, User Features and Post Features were considered. In Approach 02, User Features, Post Features and User History Features were considered. In Approach 03, User Features, Post Features, User History Features and Image Technical Features were considered, and in Approach 04, all the collected features, including Image Content Features, were considered. Table I represents these approaches with its respective features.

For each approach, Linear Regression, Ridge, Lasso, Elastic Net, Decision Tree Regressor, Random Forest Regressor, Gradient Boosting Regressor, K Neighbors Regressor and MLP Regressor models were trained. The output of the prediction is the Virality Rate which is calculated by the sum of likes and comments divided by the number of followers. Higher the Virality Rate, more likely the post (image) to go viral.

TABLE I. TABLE OF APPROACHES WITH ITS FEATURES

Approach	Features
Approach 01.a.	UF + PF (without sentiment_score)
Approach 01.b.	UF + PF (with sentiment_score)
Approach 02	UF + PF + UHF
Approach 03	UF + PF + UHF + ITF
Approach 04	UF + PF + UHF + ITF + ICF

And for each approach, the following regression evaluation metrics were calculated using scikit-learn functions to select the most suitable model.

- R-squared (R2)
- Mean Absolute Percentage Error (MAPE)
- Mean Absolute Error (MAE)
- Median Absolute Error (MedAE)

The values of the above metrics relevant to each regression model will be discussed in the Results section.

IV. RESULTS

In this section, the values of the evaluation metrics of each Machine Learning regression model are compared to predict Virality Rate. And in the Table II, the evaluation metric values of the three most suitable Regression models for each approach are shown. The Decision Tree Regressor (DTR), Random Forest Regressor (RFR) and Gradient Boosting Regressor (GBR) were found to be the top three models with the highest R-squared (R2) and the lowest Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE) and Median Absolute Error (MedAE).

TABLE II. SUMMARY OF THE RESULTS OF THE APPROACHES

Approach	Model	MAPE	MAE	MedAE	R2
Approach 1: UF+PF	DTR	0.7707	0.0173	0.0079	0.6879
	RFR	0.6362	0.0144	0.0077	0.7634
	GBR	1.6234	0.0262	0.0190	0.3757
Approach 2: UF+PF+UHF	DTR	0.1989	0.0075	0.0011	0.8243
	RFR	0.1639	0.0062	0.0014	0.8755
	GBR	0.4106	0.0115	0.0058	0.8323
Approach 3: UF +PF+UHF +ITF	DTR	0.2633	0.0096	0.0026	0.7804
	RFR	0.2098	0.0077	0.0025	0.8552
	GBR	0.4138	0.0115	0.0059	0.8317
Approach 4: UF+PF+UHF +ITF+ ICF	DTR	0.2746	0.0097	0.0028	0.7923
	RFR	0.2147	0.0079	0.0027	0.8553
	GBR	0.4112	0.0115	0.0059	0.8344

The results showed that the Random Forest Regressor shows the best performance. It has the lowest MAE, MedAE, and the highest R2, which suggests a good balance between accuracy and fit. Also it can be observed that Approach 2 shows the lowest MAPE of 0.1639 and the highest R2 of 0.8755.

Despite achieving satisfactory accuracy, additional techniques were implemented to enhance the model's reliability, assess its performance, and evaluate its generalization. Cross-validation [13], specifically the k-fold cross-validation technique [14], played a crucial role in this situation. After performing cross-validation in all four proposed approaches, it was a comfort to see almost all the MAPE, MedAE, and MAE values of each and every model have decreased, and the R2 values of each and every model have increased. Just as earlier, Random Forest performs best with a 0.15 MAPE in Approach 02 and a 0.20 MAPE in Approach 03. This can be said quite a good value when it comes to practical purposes. Hence the results after cross-validation can be observed in the Table III.

TABLE III. SUMMARY OF THE RESULTS OF THE APPROACHES AFTER CROSS-VALIDATION

Approach	Model	MAPE	MAE	MedAE	R2
Approach 1: UF + PF	DTR	0.6084	0.0172	0.0081	0.6414
	RFR	0.5555	0.0139	0.0075	0.8019
	GBR	1.5538	0.0260	0.0192	0.4320
Approach 2: UF+PF+UHF	DTR	0.1718	0.0069	0.0011	0.8554
	RFR	0.1515	0.0056	0.0013	0.9238
	GBR	0.4043	0.0110	0.0060	0.8654
Approach 3: UF+PF+UHF + ITF	DTR	0.2298	0.0090	0.0021	0.8459
	RFR	0.1981	0.0070	0.0025	0.9112
	GBR	0.4085	0.0111	0.0059	0.8677
Approach 4: UF+PF+UHF +ITF+ ICF	DTR	0.2330	0.0091	0.0023	0.8333
	RFR	0.2029	0.0071	0.0026	0.9070
	GBR	0.4057	0.0111	0.0060	0.8676

V. DISCUSSION

The differences between the actual values and the predicted values by the Random Forest Regressor model of each approach can be observed in the Fig. 2, Fig. 3, Fig. 4 and Fig. 5. The figures show that the differences between the actual values and the predicted values have been drastically reduced in the Approach 2 and thereafter.

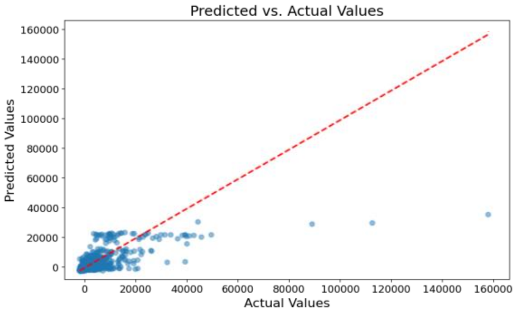

Fig. 2. *Predicted values vs Actual values of of the Approach 01*

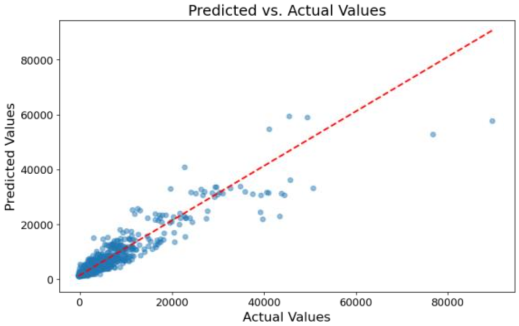

Fig. 3. *Predicted values vs Actual values of of the Approach 02*

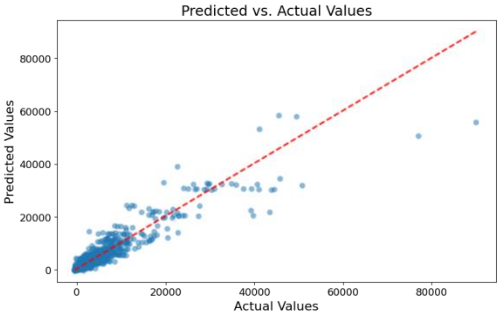

Fig. 4. *Predicted values vs Actual values of of the Approach 03*

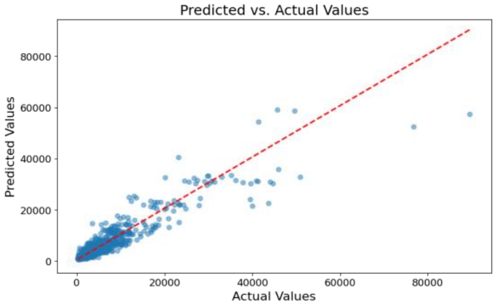

Fig. 5. *Predicted values vs Actual values of of the Approach 04*

The Random Forest and correlation-based importance graphs shown in the Fig. 6 and Fig. 7 reveal that the most crucial features for a post to go viral on Instagram are the User History Features. This explicitly says that the average number of likes, comments, ratio between the number of likes and followers, and ratio between the number of comments and followers affect the virality of a post. This further implies that influencers (users with a higher average of likes) have a higher tendency to go viral than regular people. Sentiment score is also important for virality, as the positivity of a caption or the use of emoji can increase user engagement. High hashtag scores and counts can increase virality, as popular hashtags reach more followers and users who follow or search for the post. Technical features of an image, such as high resolution, clarity, and colorfulness, can also increase virality. The correlation graph shows that the number of followings negatively impacts post Virality Rate, as a high number may make the account appear spam. Posting time, posting day, and number of mentions also slightly affect Virality Rate.

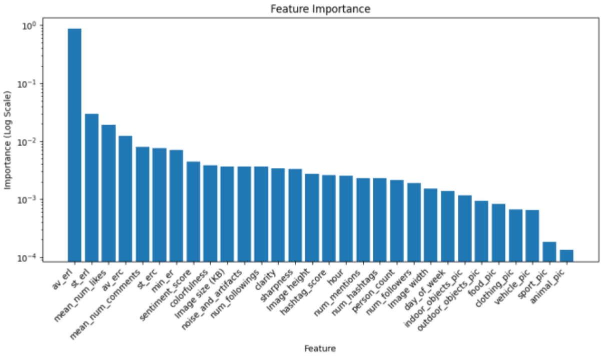

Fig. 6. *Feature Importance from Random Forest Regression Model*

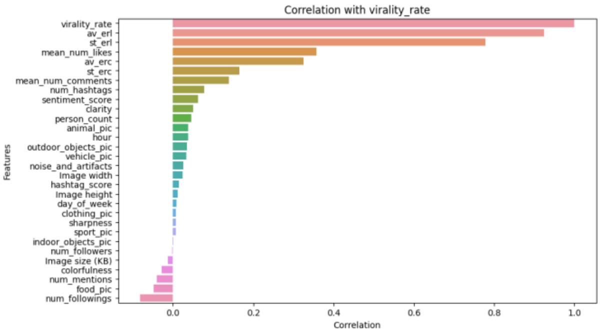

Fig. 7. *Feature Importance based on Correlation*

The Fig. 8 and Fig. 9 represent the bar charts with value importance of the posting day and time. From this research, it can be stated that a comparatively high Virality Rate can be achieved by posting on Saturdays, Wednesdays or Thursdays. And the suggested time ranges are 15:00 to 18:00, 18:00 to 21:00 and 06:00 to 09:00. This can be considered as a fair measure of time since most of the Instagram users are at work from 09:00 to 16:00. Excluding these hours and posting in the above mentioned times would increase your Virality Rate.

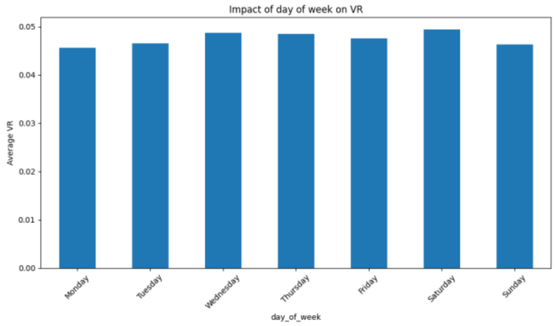

Fig. 8. *Impact on Virality Rate by the day of the week*

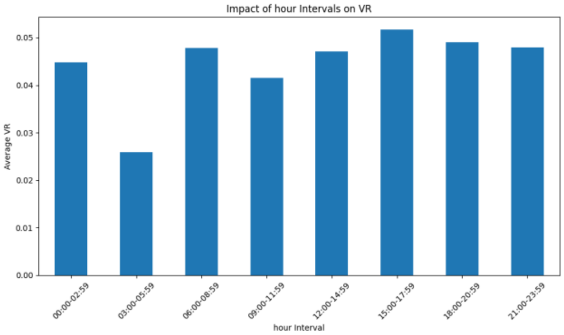

Fig. 9. *Impact on Virality Rate by the hour of the day*

VI. ACKNOWLEDGMENTS

I express gratitude to the individuals who supported me in numerous ways for their assistance and guidance in completing this research paper. I thank my supervisor, Dr. Rasika Rajapaksha, the faculty members of the Faculty of Computing and Technology, the users of Instagram, the authors of articles, research papers, and publications, and my family and friends for their motivation and belief throughout.

VII. CONCLUSION

This study aims to identify factors affecting Instagram post virality using a global dataset. Key features considered include user features, hashtag score, sentiment score of captions, image technical features, image content, and user history features. The virality analysis uses a metric called 'Virality Rate' to determine the likelihood of a post going viral.

Key features for raising Virality Rate include UHF, sentiment score, and ITF, particularly sharpness and clarity. The study also considers person count, upload day, and time. The Random Forest Regression model has the best R2 and MAPE, with the best results after considering UHF and PF. When adding ICF and ITF, the MAPE value slightly increased and the R2 value slightly dropped.

The accuracy of the model is sufficient for practical use and has a considerable rise compared to previous studies. This study will be beneficial not only for regular users but also for content creators, influencers, and marketers.

REFERENCES

[1] "Instaloader — Download Instagram Photos and Metadata." https://instaloader.github.io/ (accessed Aug. 04, 2023).

[2] M. Viola, L. Brunelli, and G. A. Susto, "Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach".

[3] Z. Zhang, T. Chen, Z. Zhou, J. Li, and J. Luo, "How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention." arXiv, Jan. 19, 2019. Accessed: Aug. 19, 2023. [Online].

[4] "Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features," IAJIT, vol. 18, no. 1, Dec. 2020, doi: 10.34028/iajit/18/1/10.

[5] M. Meghawat, S. Yadav, and D. Mahata, "A Multimodal Approach to Predict Social Media Popularity".

[6] K. Cakmak et al., "The Causal Determinants of Popularity in Instagram".

[7] C. Dugué, "Predicting the number of likes on Instagram," Medium, May 14, 2017. https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203 (accessed Aug. 03, 2023).

[8] "Top 100 HashTags on Instagram - Top-Hashtags.com." https://top-hashtags.com/instagram/ (accessed Aug. 04, 2023).

[9] "NLTK Sentiment Analysis Tutorial: Text Mining & Analysis in Python | DataCamp." https://www.datacamp.com/tutorial/text-analytics-beginners-nltk (accessed Aug. 04, 2023).

[10] "SENTIMENTAL ANALYSIS USING VADER. interpretation and classification of… | by Aditya Beri | Towards Data Science." https://towardsdatascience.com/sentimental-analysis-using-vader-a3415fef7664 (accessed Aug. 04, 2023).

[11] "The impact of image-based posts on social media insight." https://www.marketingweek.com/effect-image-based-posts-social-media-insight/ (accessed Aug. 05, 2023).

[12] "7 Proven Reasons to Use Visual Content in Social Media." https://meghanmonaghan.com/5-reasons-use-visual-content-social-media/ (accessed Aug. 06, 2023).

[13] "5 Reasons why you should use Cross-Validation in your Data Science Projects'' https://towardsdatascience.com/5-reasons-why-you-should-use-cross-validation-in-your-data-science-project-8163311a1e79 (accessed Aug. 06, 2023).

[14] "What are the advantages and disadvantages of using k-fold cross-validation for predictive analytics?" https://www.linkedin.com/advice/0/what-advantages-disadvantages-using-k-fold (accessed Aug. 06, 2023).