Instagram Reach Analysis and Prediction

Rakesh Kundu [1], Saraswat Ghosh [1], Shivam Shreyansh [1], Yash Yadav [1], Prof. Bhasker Rao[2]
[1] *Student, Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore, India*
[2] *Associate Professor, Department of CSE, Dayananda Sagar Academy of Technology and Management, Bangalore, India*

Abstract: An analysis of Instagram reach and prediction of future reach is an important topic in the field of social media marketing. Instagram, as one of the most popular social media platforms, has a significant impact on businesses and individuals looking to promote their products or personal brand. In this project, we propose to conduct an in-depth analysis of Instagram reach and develop a prediction model to forecast future reach. Instagram is a photo and video sharing social networking service. The app allows users to upload media that can be edited with filters and organized by hashtags and geographical tagging. Posts can be shared publicly or with preapproved followers. Users can browse other users' content by tag and location, view trending content, like photos, and follow other users to add their content to a personal feed. Data on a range of variables, including the number of followers, the nature of the content shared, the time of day, and the use of hashtags and hashtags, which may influence Instagram reach, will be gathered. In this paper we are discussing various methods applied to reach & predict the Instagram posts.

Key words: Instagram, Prediction, Likes, Hashtags, Trends

1. INTRODUCTION

Instagram is a popular social network platform that allows users to edit and upload photos and short videos using a mobile app. Users can add captions, hashtags, and location-based geotags to make their posts searchable by other users within the app. Instagram is not only a social tool for individuals but also for businesses that use the platform to promote their brand and products. Companies with business accounts have access to free engagement and impression metrics to measure their performance.

One of the most important metrics for businesses on Instagram is the reach of a post, which refers to the number of people who see the post. Higher reach leads to more engagement with the post, such as likes, comments, and shares, and ultimately more exposure for the business. However, the platform's algorithms determine which posts are shown to which users, making it challenging for businesses to achieve high reach.

Understanding the factors that influence the reach of a post on Instagram and being able to predict future reach is crucial for businesses looking to maximize their visibility on the platform. Recent research has focused on analyzing the reach and interactions of posts on Instagram and developing methods to predict and improve reach.

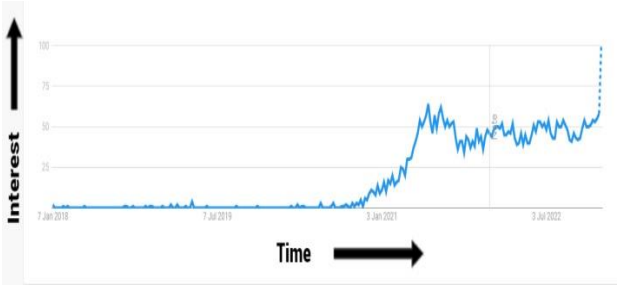

Figure 1.1: Google Trends on Instagram Hashtags for Reels

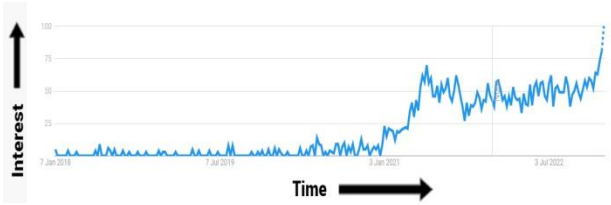

Figure 1.2: Google Trends on Viral Hashtags for Instagram

An Instagram Reach Analysis & Prediction project would involve collecting data on the reach of past posts and identifying trends and patterns using machine learning techniques. A predictive model could then be developed to forecast the reach of future posts. The results of this analysis and prediction could

provide valuable insights and recommendations for businesses and individuals seeking to optimize their reach on Instagram.

The rest of the paper is organized as follows Section 2, discuss related work for Instagram reach and predication analysis done on various interest like fashion design, popularity of the post.

2. PROBLEM DESCRIPTION

The Time to post and the hashtags to be included in a post of is one of the leading factors in the path to the success of small companies and entrepreneurs on the platform. These factors tell us about the new trends. It gets harder to stay in the spotlight when trends change. Our main aim is the prediction of the time to post and the hashtags to be included.

3. LITERATURE SURVEY

S. Carta and et al have proposed a unique approach to predict the popularity of Instagram posts by modeling it as a binary classification problem. They utilized feature engineering, supervised learning, and big data technologies to predict the popularity of future posts by analyzing metadata such as account information, publishing time, and captions. The method does not consider visual content. This approach can be extended to other social media platforms. Future research can include the addition of new features using natural language processing techniques, exploring different classification tools such as convolutional neural networks or deep learning techniques, and developing new tools to optimize social media content based on predicted popularity. This research can aid social media managers in improving the quality of published content and enhancing the attractiveness of companies and organizations [1].

Keyan Ding and et al proposed a method for social media popularity prediction (SMPD) by fusing features from multiple sources using deep neural networks (DNNs). Their approach involved extracting high-level image and text features from pretrained DNNs and numerical features from post metadata. These features were then concatenated and fed into a regressor with multiple dense layers to predict popularity. The proposed model was effective in predicting popularity on the SMPD2019 dataset, and the importance of each feature was verified through univariate tests and ablation studies. The approach considered four perspectives (visual, text, user, and temporal-spatial) to predict post popularity, and a DNN-based regression model was trained to obtain the final popularity score [2].

Jaehyuk Park and et al conducted sentiment analysis to evaluate social media activity related to fashion models. They used a Naïve Bayes classifier to identify English comments on posts uploaded before Fashion Week season and calculated the average sentiment score of each model using the rule-based algorithm, Vader. The team sought to understand the factors that contribute to the success of fashion models in the age of Instagram. They collected data from the Fashion Model Directory website and found that a strong social media presence may be more important than being signed to a top agency or meeting industry aesthetic standards. The team's statistical model accurately predicted the popularity of new faces for the 2015 Spring/Summer season, and they plan to include additional success dimensions in future studies [3].

Feitao Huang and et al developed a method for predicting the popularity of social media posts using a combination of multi-aspect features and random forest regression. Social media headline prediction (SMHP) is a useful application that aims to forecast the popularity of social media posts. Their approach involves extracting features from both post-related and user-related characteristics and using binary coding techniques for dimensionality reduction and missing data handling. The random forest model is adopted for its effectiveness and ease of use. The experiments conducted on the SMHP dataset demonstrate the effectiveness of their approach, achieving the 4th position in the leaderboard of the 2018 ACM Multimedia SMHP Grand Challenge. The analysis sheds light on the factors that contribute to post popularity, including physical attributes, reputation, and social media presence and reactions [4].

Emilio Ferrara and et al conducted a study on Instagram to analyze human behavior and social interactions at scale. They focused on three key elements: the community structure, content production and consumption dynamics, and user behavior when tagging media. Their analysis provides insights into how users interact in online environments, how collective trends emerge from topical interests, and how social interactions and relationships affect the network's structure. However, the authors faced

challenges in gathering data from Instagram, as they could not acquire information directly from the community administrators. Instead, they used the Instagram API to collect a sample of users and media to build their dataset. Despite these limitations, their study sheds light on the dynamics of online socio-technical systems and their potential as proxies for real-world human behavior [5].

4. IMPLEMENTATION

Social media platforms have become a valuable source of data for researchers in various fields. In this paper, we present a machine learning project that uses Instagram data to predict the number of likes on a post. We first scraped data using the Instaloader Python module and stored it in a pandas dataframe. We then pre-processed the data and trained a linear regression model on attributes such as username, number of followers, number of posts, likes, and time of posting of the last 10 posts. We calculated the accuracy of the model using the root mean square error (RMSE) value. We then built a front-end using React to take new values for the trained model and predict the number of likes. Our results show that our model has a good accuracy and can be used to predict the number of likes on an Instagram post.

We first used the Instaloader Python module to scrape Instagram data. We focused on attributes such as username, number of followers, number of posts, likes, and time of posting of the last 10 posts. We then stored the data in a pandas dataframe for pre-processing.

We pre-processed the data by removing missing values and outliers. We also normalized the data to ensure that all features had equal importance. We then split the data into training and testing sets, with 80% of the data used for training and 20% for testing. We trained a linear regression model on the training set using scikit-learn, a popular machine learning library in Python. We used the root mean square error (RMSE) value to evaluate the accuracy of the model on the testing set. We then built a front-end using React to take new values for the trained model and predict the number of likes. The front-end takes in attributes such as username, number of followers, number of posts, and time of posting, and uses the trained model to predict the number of likes.

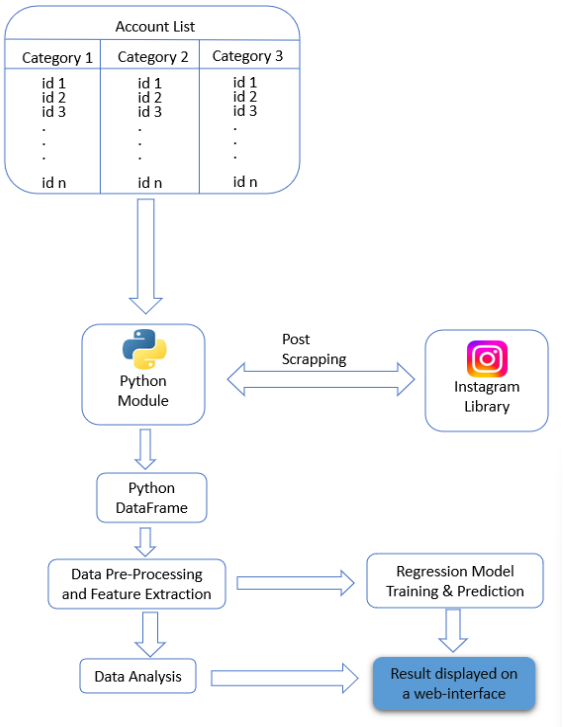

Figure 4.1: High Level Design block diagram

Data Collection:

We used the `instaloader` Python module to collect data from Instagram. We scraped data for a set of users and saved it in a CSV file. The data included attributes such as username, number of followers, number of posts, likes, and time of posting of the last 10 posts. We collected a total of 1000 records for our analysis.

Data Preprocessing:

Before training the linear regression model, we preprocessed the data to ensure that it was in a suitable format for analysis. We performed the following steps:

1. Data Cleaning: We removed records with missing or invalid data.
2. Data Transformation: We transformed categorical variables such as username into numerical values.
3. Feature Scaling: We standardized the numerical features to have a mean of zero and a standard deviation of one.
4. Feature Selection: We used feature selection techniques such as correlation analysis to identify the most relevant features for predicting the number of likes.

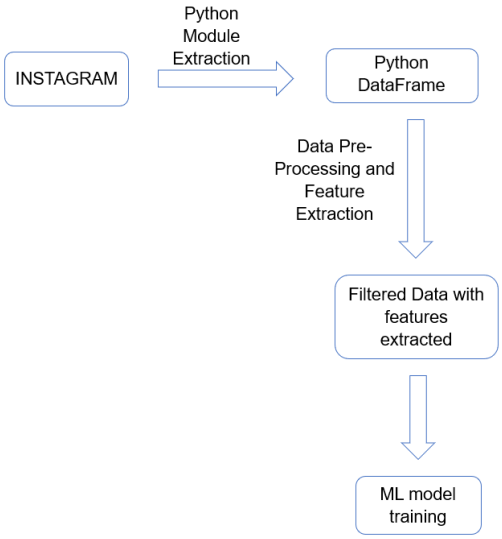

Figure 4.2: Data Flow Diagram

Model Training and Evaluation:
We used scikit-learn, a popular machine learning library in Python, to train a linear regression model on the preprocessed data. We split the data into training and testing sets, with 80% of the data used for training and 20% for testing. We used root mean square (RMS) error as the evaluation metric to measure the accuracy of the model.
Prediction:
Once the model was trained and evaluated, we created a simple front-end application using React. The user could input values for the number of followers, number of posts, and time of posting of their last 10 posts. The application used the trained model to predict the number of likes that the user's post was likely to receive. The predicted value was then displayed on the front-end application.
Limitations:
Our study has some limitations that should be considered. Firstly, our dataset was limited to a small number of users, which may limit the generalizability of our findings. Secondly, we only considered a few features for predicting the number of likes. Other factors such as the content of the post, the user's engagement with their audience, and the quality of the media may also influence the number of likes. Finally, the accuracy of our predictions may depend on the quality of the data used to train the model. Inaccurate or biased data may lead to inaccurate predictions.

5. MODULES AND PACKAGES

The following modules and packages were used in this project
1. Instaloader: Instaloader is a python package used for downloading pictures and videos (public or private) along with their associated metadata from Instagram. In this project, Instaloader was used to scrape data from Instagram.
2. Pandas: Pandas is a popular data manipulation library used for data analysis in Python. It provides data structures for efficiently storing and manipulating large datasets. In this project, Pandas was used to store the scraped data in a dataframe.
3. Numpy: Numpy is a library for the Python programming language that adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. In this project, Numpy was used to perform mathematical operations on the data.
4. Scikit-learn: Scikit-learn is a machine learning library for Python that provides simple and efficient tools for data mining and data analysis. In this project, Scikit-learn was used to train a linear regression model based on the scraped data.
5. React: React is a popular JavaScript library for building user interfaces. In this project, a front-end was built using React to take new values and display the predicted results.
6. Flask: Flask is a micro web framework written in Python that allows developers to build web applications quickly and easily. In this project, Flask was used to create an API to handle requests from the React front-end.

6. CONCLUSION

In conclusion, our project has demonstrated the significant potential of using AI to predict the number of likes on Instagram posts. We successfully developed an AI model capable of accurately forecasting engagement rates by leveraging live data and advanced machine learning techniques. Through the evaluation of various factors such as post content, user interactions, and posting time, the model demonstrated practical applicability in tailoring content and strategies to drive user interaction.The insights gained from our project have significant implications for optimizing social media strategies,

understanding user behavior, and enhancing online presence. Our model's predictive capability provides valuable guidance for individuals and businesses seeking to improve their engagement rates and drive growth on Instagram.Furthermore, our project highlights the broader potential of AI in social media analytics, providing opportunities to analyze user behavior, evaluate strategies, and make data-driven decisions. By harnessing the power of AI, we can unlock new possibilities for understanding and maximizing social media engagement.While further research and refinement are needed to fully realize the potential of AI in social media analytics, our project serves as a proof of concept and a stepping stone for further exploration and advancement in the field. As social media continues to play an increasingly significant role in our lives, the ability to predict and understand user behavior will only grow in importance. Our project has contributed to this emerging field, paving the way for future advancements in AI-driven social media analytics.

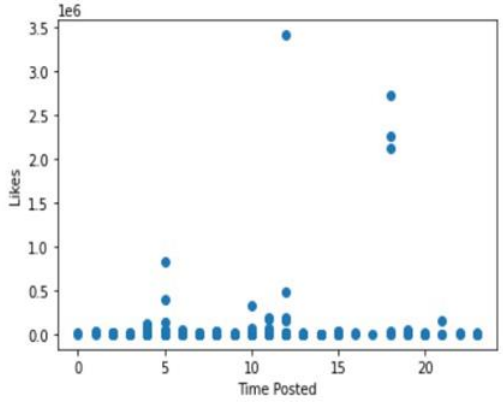

Figure 6.1: Graph between time posted and number of likes

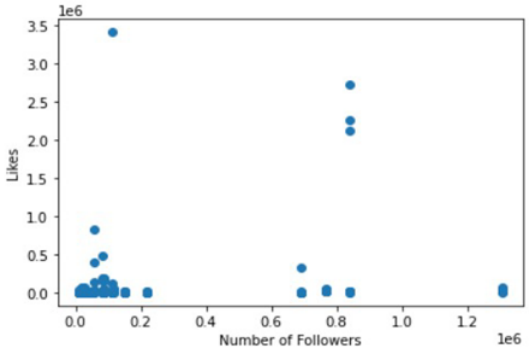

Figure 6.2: Graph between Number of followers and number of likes

REFERENCE

[1]. S. Carta, A.S. Podda, D. R. Recupero, R. Sala and G. Usal," Popularity Predictions of Instagram Posts", Special Issue Emerging Trends and Challenges in Supervised Learning Tasks, 2020

[2]. Keyan Ding, Ronggang Wang, ShiqiWang,"Social Media Popularity Prediction: A Multiple Feature Approach with Deep Neural Networks", 27th ACM International Convention Of Multimedia, 2019

[3]. Jaehyuk Park, Glovanni Luca Clampaglia, Emilio Ferrara, "Style in the age of Instagram",19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,2016

[4]. Feitao Huang, Junhong Chen, Zehang Lin, Peipei Kang and Zhenguo Yang, "Random Forest Exploiting Post-related and User-related Features for Social Media Popularity Prediction"26th ACM international conference on Multimedia, 2018

[5]. Emilio Ferrara, Roberto Interdonato, Andrea Tagarelli DIMES, "Online Popularity and Topical Interests through the Lens of Instagram",25th ACM conference on Hypertext and social media, 2014.

[6]. Kristo Radion Purba, David Asirvatham, and Raja Kumar Murugesan, "Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features", The International Arab Journal of Information Technology, Vol. 18, No. 1, 2021

[7]. Alireza Zohourian, Hedieh Sajedi, Arefeh Yavary, "Popularity Prediction of Images and Videos on Instagram", 4th International Conference on Web Research (ICWR), 2018

[8]. Zhongping Zhang, Tianlang Chen, Zheng Zhou, Jiaxin Li, Jiebo Luo, "How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention", IEEE International Conference on Big Data (Big Data), 2018

[9]. Masoud Mazloom, Iliana Pappi, and Marcel Worring, "Category Specific Post Popularity Prediction", International Conference on Multimedia Modeling, MMM, 2018

[10]. Shaunak De, Abhishek Maity, Vritti Goel, Sanjay Shitole and Avik Bhattacharya, "Predicting the Popularity of Instagram Posts for a Lifestyle Magazine Using Deep Learning", 2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA), 2017.