Instagram User Popularity Predictor

RAVI GUSAIN[1], SAKSHAM PATHAK[2]

[1, 2] *Department of Information Technology, Maharaja Agrasen Institute of Technology, Delhi, India*

Abstract- Instagram is photo and video sharing platform having a huge influence on the reach of a person or a brand worldwide. Popularity prediction is advantageous for influencers as well as organisations that make revenue through advertising or product placements. The main aim of this study is to find out the relationship between the popularity of an Instagram post with the image content posted and it's metadata like time, day, number of hashtags, number of comments, etc. This research was conducted using data scraped from various active Instagram accounts and applying regression models to gather relevant metrics to predict the likelihood of the user's posts being well received. Two regression techniques were tested. The Linear Regression model successfully predicted the number of likes with an MSE of 953.76, whereas the XGB Regression model had a MSE of 2876.17. Rather than viewing just the follower count for the prediction, the post's metadata was also a major contributor.

I. INTRODUCTION

Instagram reached two billion active users in Q3 2021 and has continued to grow at a steady pace. It is on track to reach 2.5 billion by 2023 [1]. These many users means an infinite potential for advertising and product placements that can increase a business' potential exponentially. In order to find the best content creator/influencer to advertise through, there must be some method to analyse their popularity other than just looking at the number of followers and likes each post generates. This will allow for new marketing strategies for brands and businesses.

Instagram provides a various array of data on a post and the user. This data can be combined to form multiple metrics to analyse the popularity of the given post and to further it can be used to predict the popularity of a future post the user may want to make.

The goal of this study is to be able to accurately anticipate the likelihood of a post going viral, hence determining whether the user is popular enough to be invested into. The study will also determine whether a particular metric is actually useful in viewing the user's popularity, for example whether the number of followers of the user actually mean more social media presence, or whether the number of comments have any influence on popularity.

II. LITERATURE REVIEW

There have been previous studies on the topic of predicting user popularity or even the number of likes they may get on an Instagram post. Almost all the prior research is done by looking at the post metrics directly provided by the Instagram website such as number of prior posts, followers and following, or likes on previous posts. More data such as image content, associated hashtags, and description has also been used to gather more data by using image analysis libraries or sentiment analysis algorithms.

The first step of gathering dataset is common in all the predecessors – Using web scraping/automation tools to read data off of Instagram [2, 3, 4, 8]. Their reasoning for preferring this over using the pre-existing Instagram API [5] is that the API has a limited number of requests per day which is too slow to gather reasonable amount of data.

Some of the common details taken were:
1. Post caption of each post
2. Thumbnail image of each post
3. Hashtags used in each post
4. Topic assigned for each post
5. Time of post during a day
6. Length of post caption
7. Hashtag count in a post

According to one paper websites such as "Heepsy"[10], "HypeAuditor" [11], "Coobis" [12] were used. There are AI-powered Instagram data collection platforms that directly provide the required metadata for the dataset.

For basic analysis, creating graphs and data analysis the popularity of a post is defined by the like-follower ratio LTF [4]. With these graphs that are plotted using the metadata mentioned above a general hypothesis can be formed and a set of variables to look out for can be created that will be further used in training the models.

As for the algorithms and methodologies used, for training the text model cbow [3] ("continuous-bag-of-words") is used as well as a Neural network model architecture [2, 3]. As for image content, they are inputted into a pre-trained CNN as done in [2, 3, 9]. Variables used to analyse the images were image aesthetic and image quality which were inputted into different algorithms like Google's Neural Image Assessment (NIMA) [15]or pic2vec [16].

III. METHODOLOGY

The methodology will be split into four phases, i.e. Building the dataset, Scraping user data, Dataset Analysis and developing the prediction model.

A. *Building the dataset*

The first step in building the dataset has to do with collecting a list of Instagram influencers. Upon searching on Google for the top influencers on Instagram we came across a website "Qorus" [13] which contains a list of top influencers along with their ids. We have to scan multiple pages to get a good sample size of the influencers. This website along with basic information like Instagram id and number of followers also gives the average engagement (Average number of Likes), though this much information is not enough. Hence we have to use the Instagram API or build our own web scraper in order to look-up the profiles and their posts for gathering more data.

B. *Scraper*

The next step is building the scraper in order to extract the user data from Instagram. The Instagram API has not been used since with a limit of 60 requests per hour it is not practical to wait that long. The scraper is coded in Selenium; a website testing tool. This scraper scans the latest posts of a user, then opens each post to retrieve more granular information related to each image.

C. *Data Cleaning*

After running the selenium script, the following data is obtained containing all the metadata form the post as well as the image.

The data obtained is then analysed on the basis of the following fields:

- Followers
- Following
- Likes
- Comments
- Number of tags
- List of tags
- Date
- Day of week
- Type of post
- Number of users
- Month hour
- Caption length

D. *Visualising the variables:*

By plotting histograms we are able to visualize the data for each individual variable. According to the histogram (Fig 3.1) majority of the Instagram users taken have following of about 1 lakh (10^5) with few reaching up to 11 lakhs. Meaning the users can be taken as an influencer and not an average user of Instagram.

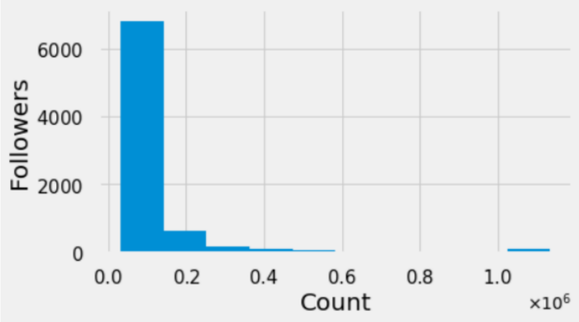

Fig 3.1 Count of number of followers

Another common theme that was observed is that it was not necessary that high number of followers would mean a high number of likes per post. As seen below (Fig 3.2) the average number of likes does not vary too much. This means an influencer's follower count alone cannot be taken as the user's popularity measure.

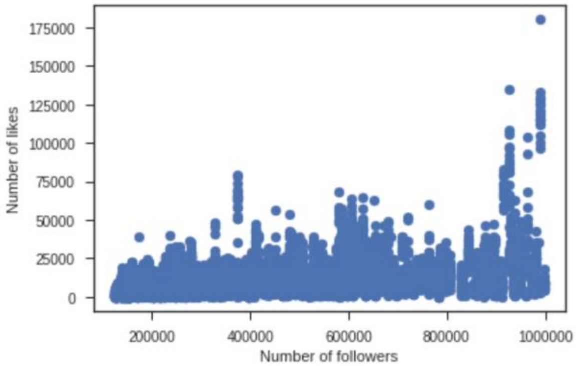

Fig 3.2 Number of followers to Number of likes

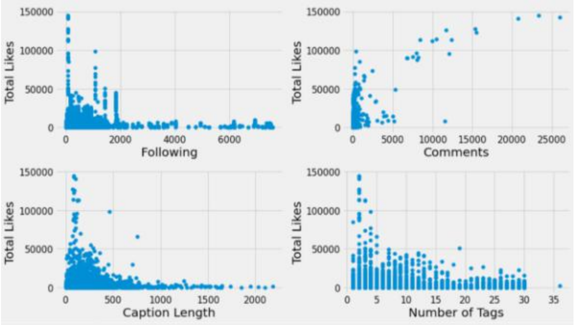

Fig 3.3 Total Likes to Quantitative Predictors Scatter Plot

The majority of the connections in Fig3.3 seem to be non-linear. The scatterplots for followers (Fig 3.2) and comments (Fig 3.3 top-right) show a strong correlation between those factors and overall likes. The scatterplots for caption length, following, users, and tags indicate a negative correlation between these predictors and the overall amount of likes.

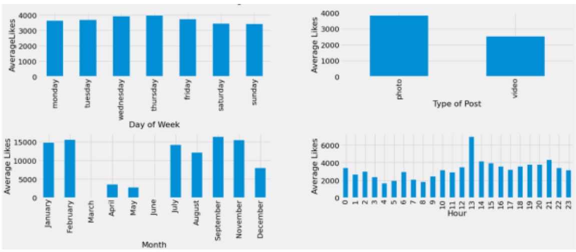

Fig 3.4 Bar Charts for Categorical Variables to Likes

It is clear from the top-left Bar Chart (Fig 3.4) that the day of the week does not have much impact on the number of likes received. The higher number of likes on posts with images as shown in top-right suggests people prefer liking images over videos. Though the bottom-left chart may suggest low number of likes between March and June, it can also be attributed to the fact that there was not much data in the dataset for those months. The time graph at bottom-right shows the preferable time is at 13:00 hours (1:00 PM).

E. Models

In order to give a prediction we will have to train a model with the variables mentioned in the previous section. We will work in the following order to get the best prediction outcome. We will first develop a base model using Linear Regression Algorithm that takes Followers over Time as its parameters for training. Then, we will add features obtained from data pre-processing through Natural Language Processing (NLP) feature extraction. We also apply an XGBoost model on the data to get a feature importance plot. The different models can be compared using two performance metrics: Root Mean Square Error (RMSE) and the R² value [17].

Linear Regression:

Linear regression analysis is used to predict the value of the Number of likes based on the value of the Follower count and time since posted.Reason for choosing Linear Regression is that the mathematical technique used in linear-regression models is straightforward and can be used to make predictions.

XGBoost:

XGBoost stands for eXtreme Gradient Boosting. All the data from the dataset is fed to the XGBRegressor which can then be compared with the actual value of the number of likes to get a score.

Since XGBoost Regression is done by taking multiple features, taking features Number of Likes and Number of Followers into account and with the following parameters:
max_depth=4, learning_rate=0.01, n_estimators=596

IV. EXPERIMENTAL RESULTS

I was observed in the previous chapter that the follower count alone of a user is not a good measure of their popularity due to there being low variation in the number of likes with change in number of followers.

Fig 4.1 plots the predicted as well as the actual values of Number of Likes versues each Post in the dataset. The main accuracy measure used was Mean Square Error with the Linear Regression model having a Mean Squared Error of 953.76.

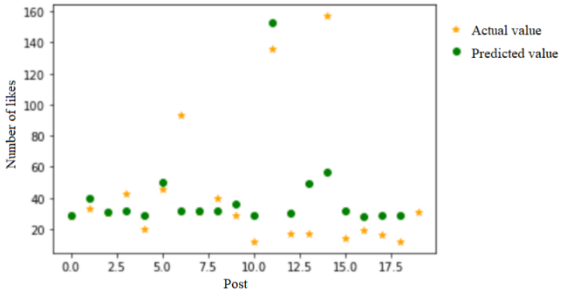

Fig 4.1 Graphical representation of Actual and Predicted Result using Linear Regression

Fig 4.2 shows the deviation between the Actual and Predicted results as given by the XGB Regressor. The further away the points are from the solid line the greater will be the error. The final mean squared error obtained was 2876.17 with a learning rate of 0.035.

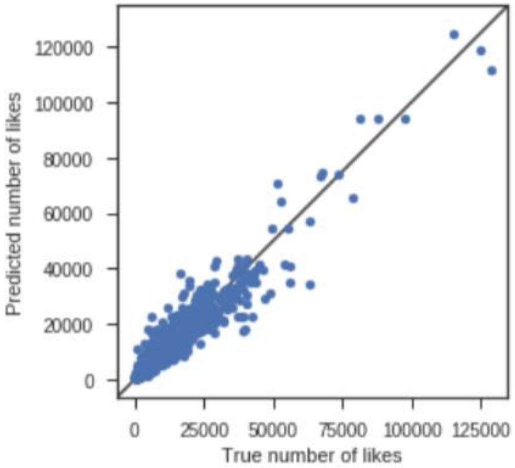

Fig 4.2 Predicted number of likes to true number of likes using XGB Regressor

Using XGB's plot_importance function as depicted in Fig 4.3 below, we can see that the average number of likes had a huge impact on the on the outcome of the XGBoost model. The higher F score of 4084 for averageLikes indicates so. The F score is based on the frequency with which a variable is chosen for splitting, weighted by the squared improvement to the model brought about by each split, and averaged over all trees.

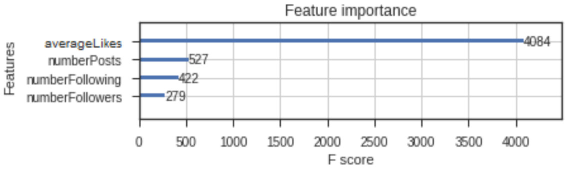

Fig 4.3 Feature importance graph

CONCLUSIONS AND FUTURE SCOPE

Using a set of variables obtained from user metadata, posts, hashtags, image evaluation, and user history, we have analysed and predicted Instagram's popularity trends in this paper. With all features, prediction accuracy can be as high as 91.74. This accuracy is significant as compared to earlier research and is sufficient for practical application. Regular users can utilise the popularity trend data to their advantage when figuring out how to increase likes. Users who are in the business world can also gain from identifying influencers for brand marketing.

However, threat to validity of this research work is that the sample size of the data set is limited with only a hundred or so influencers being taken into account. It is common knowledge that the bigger the dataset the better trained the model will be. Hence in a future research, it will be better to use a larger dataset.

In future this research, including the comments for each post will help in increasing the accuracy level of the proposed model. Using text/sentiment analysis for each comment can be a great way to review the positive/negative feedback to the post. This feature will also allow for filtering of less meaningful spam posts that may have fabricated their number of likes or followers.

REFERENCES

[1] Business of Apps Analysis, https://www.businessofapps.com/data/instagram-statistics/ Acessed on 8 October 2022

[2] https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203; Accessed on 24 October, 2022

[3] Joel CanteroPriego "Predicting the number of likes on Instagram with TensorFlow", UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) BarcelonaTech, October 26, 2020

[4] Crystal J. Qian, Jonathan D. Tang, Matthew A. Penza, Christopher M. Ferri "Instagram Popularity Prediction via Neural Networks and Regression Analysis"

[5] KristoRadionPurba, David Asirvatham, and Raja Kumar Murugesan "Instagram Post Popularity Trend Analysis and Prediction using Hashtag, Image Assessment, and User History Features", School of Computer Science and Engineering, Taylor's University, Malaysia, 1, January 2021

[6] KristoRadionPurba, David Asirvatham, Raja Kumar Murugesan, "Analysis and Prediction of Instagram Users Popularity using Regression Techniques based on Metadata, Media and Hashtags Analysis", 28 March 2020.

[7] Salvatore Carta , Alessandro Sebastian Podda * , Diego ReforgiatoRecupero , Roberto Saia and Giovanni Usai, "Popularity Prediction of Instagram Posts", Department of Mathematics and Computer Science, University of Cagliari, 09124 Cagliari, Italy; 18 September 2020

[8] Yu-Yun Liao, "Leveraging Hashtag Networks for Multimodal Popularity Prediction of Instagram Posts", Graduate Institute of Linguistics National Taiwan University; 20-25 June 2022

[9] Massimiliano Viola , Luca Brunelli , and Gian Antonio Susto, "Instagram Images and Videos Popularity Prediction: a Deep Learning-Based Approach", Universit`adegliStudi di Padova, Padova, IT

[10] Heepsy, https://www.heepsy.com/; Accessed on 24 October, 2022

[11] Hyper Auditor, https://hypeauditor.com/; Accessed on 24 October, 2022

[12] Coobis, https://coobis.com/en/; Accessed on 24 October, 2022

[13] Qoruz, https://qoruz.com/find-influencers/top-100-instagram-influencers-india

[14] Open Source Library, https://github.com/idealo/image-quality-assessment

[15] Talebi H. and Milanfar P., "Nima: Neural Image Assessment," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3998-4011, 2018.

[16] Open Source Library, https://github.com/datarobot/pic2vec; Accessed on 24 October, 2022

[17] Casella, Georges (2002). Statistical inference (Second ed.). Pacific Grove, Calif.: Duxbury/Thomson Learning. p. 556. ISBN 9788131503942.