Analysis and Prediction of Instagram Users Popularity using Regression Techniques based on Metadata, Media and Hashtags Analysis

Kristo Radion Purba, David Asirvatham, Raja Kumar Murugesan

***Abstract*— In recent years, social media is growing at an unprecedented rate, and more people have become influencers. Understanding popularity helps ordinary users to boost popularity, and business users to choose better influencers. There were studies to predict the popularity of posted images on social media, but there was none on the user's popularity as a whole. Furthermore, existing studies have not taken hashtag analysis into consideration, one of the most useful social media feature. This research aims to create a model to predict a user's popularity, which is defined by a combination of engagement rate and followers growth. There were six machine learning regression models tested. The proposed model successfully predicted the users' popularity, with R^2 up to 0.852, using Random Forest with 10-fold cross-validation. The additional statistical analysis and features analysis results revealed factors that can boost popularity, such as actively posting and following users, completing user's metadata, and using 11 hashtags. In contrast, it was also found that having a large number of posts and following in the past will not help in growing popularity, as well as the use of popular hashtags.**

***Index Terms*—Machine Learning, Regression Analysis, Social Media, Predictive Model**

I. Introduction

Social media is a powerful tool for spreading news [1], advertisement [2], and persuading others [3]. As of July 2018, Instagram was the third most popular social media, with 1 billion active users [4]. The rise of social media causes the rise of influencers, which are users with significant number of followers [5] or strong identity [6].

Despite being the third in terms of active users count, Instagram is the most popular platform for brand marketing [7]. Social media marketing has more impact compared to traditional counterpart. Quick rise of visibility [8], cheaper [9], and deeper engagement [10] are among the main benefits of the former.

The emergence of new influencers, further supported by the demand from business owners, have led many people to try to increase popularity [11]. However, determining factors that contribute to popularity is often a difficult task to breakdown. For example, a usual number of followers growth rate on Instagram is 5-7.5% per month [12], however, an unpredictable growth can sometimes happen in just two months [13].

Recognizing the current trend in popularity, this study aims to analyze user data and create models to predict user popularity on Instagram. Popularity is defined by a combination of followers growth per month and engagement rate (from likes and comments). This research can benefit both parties, i.e. to help regular users in understanding popularity, and business users in picking influencers.

The following questions are addressed in this research, i.e. (R1) Which regression model is the most accurate in predicting popularity? (R2) What is the contribution of each feature to a user's popularity? (R3) What is the effect of using popular hashtags and unique hashtags in popularity?

This study used features collected from user's metadata, media data, hashtag similarity and popularity. Existing studies lacked hashtag analysis, and there was no study that predicted user popularity with follower growth as the popularity measure. Most studies made predictions on post popularity, with the number of likes as the measure.

There were six machine learning regression models tested in this study, namely linear regression, multivariate adaptive regression splines (MARS), multilayer perceptron (neural network), Random Forest regressor (RF), XGBoost, support vector regression (SVR). These models were tested using Weka [14] using the provided default parameters.

II. Related Works

There were studies on popularity prediction and analysis using various features. Other than the user's metadata, the most important features were the use of hashtags [15] [16] [17] and the type of photos [16] [18].

Hashtag is a powerful tool to raise the visibility of a product [19] [20]. It is also one of the most important contributors to the popularity of the foods and beverages industry [21]. However, a quick observation revealed that some users tend to use hashtags excessively. This includes maximizing the number of hashtag (maximum of 30 on Instagram), or similar hashtags, such as #food, #instafood, #foodmalaysia, etc. This research will investigate the effect of similar hashtags, popular hashtags and hashtags count in users' popularity. To our imagination, similar hashtags only attract the same group of people, while the usage of popular hashtags will attract more likes.

Manuscript received December 13, 2019.

Kristo Radion Purba is with the School of Computing and IT, Taylor's University, Malaysia (Phone: +6016-529-2630; e-mail: kristoradionpurba@sd.taylors.edu.my).

David Asirvatham is with the School of Computing and IT, Taylor's University, Malaysia (e-mail: david.asirvatham@taylors.edu.my).

Raja Kumar Murugesan is with the School of Computing and IT, Taylor's University, Malaysia (e-mail: rajakumar.murugesan@taylors.edu.my).

Recent studies on popularity were the prediction of posts popularity, instead of user popularity, as listed in Table I. There are two category of prediction methods, namely classification (for categorical output) and regression (for numerical output). In recent studies, the most used regression models were linear regression (LR) and support vector regression (SVR). While the LR is the easiest to interpret, it's not as accurate as non-linear methods. Other popular and best performing regression models are MARS [22] [23] [24], Random Forest [25], neural network [23] [25], XGBoost [26].

III. Research Methodology

There were four phases on this study, i.e. data collection in two periods, data pre-processing, implementation results, and features analysis. The methodology is shown in Fig. 1.

A. *Data Collection*

The dataset was collected using scrapping technique from several third-party Instagram websites, in two periods, i.e. August 2019 (period 1) and September 2019 (period 2). The users were collected from the followers of 20 Instagram pages in Malaysia, i.e. 5 private universities, 5 public universities, 5 business accounts, 5 shopping malls. The period 2 was intended to compare the number of followers, following, and posts. Due to some differences in the exact time of scraping between a user in the two periods, these numbers were normalized to 30 days.

Users with zero followers in both periods were removed. Users with zero followers in period 1, but any non-zero follower in period 2, were set to 100% growth. There were only 864 such users.

B. *Data Preprocessing*

To increase the data accuracy, removal of suspicious (potentially fake) users were conducted. Three human annotators were assigned to flag suspicious accounts. If two or more annotators flagged a user, the user will be removed. After this removal, the number of users was down to 58,981 from 71,585 users. This process, though, was based on human judgement. While CAPTCHA validation [27] is more accurate, validation on a large number of users is a challenge.

In this research, no outliers were removed. Outliers are data that is significantly different from other observations [28]. Although such removal can produce better regression results, outliers were kept to capture the dynamics of the behaviors on a social network.

There were 14 features (or independent variables) used for the regression model, as shown in Table II, based on user's metadata and media. The metadata from external factors, such as the number of likes, comments, followers were used as the output. Thus, all of the input features were internal factors, which are controllable by a user.

The output (or dependent variable) is *ppl* (popularity index), which is a combination of engagement rate and followers growth rate. As they are two common popularity metrics [29] [30], they are given an equal proportion in *ppl*. Each number was normalized before summed up. The features definitions are given in Table II.

TABLE I
PREVIOUS STUDIES ON POPULARITY PREDICTION ON INSTAGRAM

Ref	Features	Method	Context	Output
[45]	Image content, metadata, text analysis	Dual attention	Popular image	Likes
[46]	Changes in likes and hashtags through time	LSTM	Top hashtags	Hashtag popularity
[16]	Hashtag, image content	LR	Image types	Likes
[18]	Metadata, image aesthetic, media data	NN, LR	Image types	Likes
[47]	Time, hashtag, media data	Deep learning	Lifestyle	Likes
[44]	Image content, sentiment, follower	SVR	Brand	Likes
[48]	Video analysis (face, color, text), views	SVR	Various videos	Video popularity

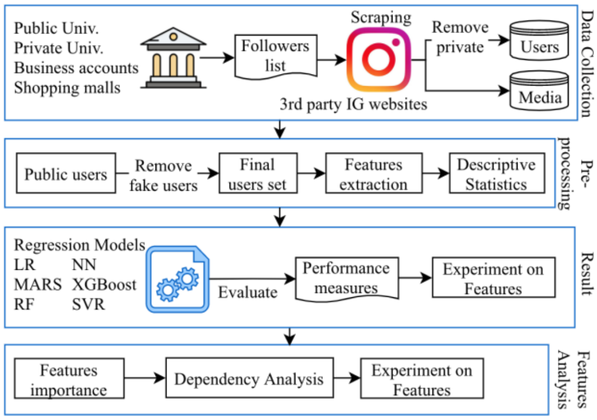


Fig. 1. Research Methodology

TABLE II
FEATURES DEFINITION

Feature	Definition	Source
pos	Number of posts, normalized to 0.00-1.00 (actual range: 0 to 2900)	d1
flg	Number of following, normalized to 0.00-1.00 (actual range: 0 to 6000)	d1
posd	Difference of number of posts in 1 month	d12
flgd	Difference of number of following in 1 month	d12
bl	Biography length (characters)	d1
pic	Profile picture (0 if none, 1 if has)	d1
link	Biography link (0 if none, 1 if has)	d1
cl	Average caption length (characters)	m1
ni	Non image percentage (percentage of posts that is video or carousel). A carousel is multiple images.	m1
lt	Average location tag availability (average of [0 of none, 1 if has location] from posts)	m1
hc	Average hashtags count	m1
pi	Average post interval (in hours)	m1
hp	Total hashtags popularity. The *hp* score of a hashtag is the total posts on Instagram that used the hashtag. The *hp* of each user was normalized to 0.00 to 1.00.	m1
hs	Hashtag similarity score, i.e. the average of the percentage of the closeness of each hashtag pair. Closeness is how often each two hashtags come together in the dataset. The score was normalized to 0.00-1.00. An example is shown in Fig. 1.	m1
Output	**Definition**	
ppl	**Popularity index, defined by: 0.5*er + 0.5*fg**	-
fg	Followers growth %, normalized to 0.00 to 1.00	d12
er	Engagement rate %, normalized to 0.00 to 1.00, i.e.: (likes + comments) / (number of media) / (followers)	m1

Source: d = Metadata, m = Media, from up to 12 latest posts
1 = Period 1, 2 = Period 2
12 = Difference between period 1 and period 2, normalized to 30 days

C. Performance Indicators

The used performance measures are R^2 for accuracy or goodness-of-fit, along with commonly used error measures [31] [32]. The details are shown in Table III.

D. Descriptive Statistics

The statistical values of each variable are presented in Table IV. The values show a dynamic behavior of users in social media. For example, the {*flgd*, *posd*} minimum and maximum values show that some users are very active, and some have these numbers decreasing in a month. The {*pos*, *flg*, *hp*, *hs*} values are normalized values.

In order to understand the differences between popular and less popular users, users are divided into nine tiers based on *ppl*. The average of each feature for each user tier is shown in Table V, and the tiers category is shown in Table VI. It can be concluded that higher *posd* and *flgd* affect popularity since higher tier users were using higher *posd* and *flgd*.

IV. Regression Results

In this section, the prediction results using six regression models and 10-fold cross-validation are presented. The results are presented in Table VII. The linear regression (LR) model analyzes the data in a linear way, unlike other methods.

The Random Forest (RF) regressor produced the best result, with an R^2 value of 0.852. This shows that 85.2% variance of the user's popularity can be explained using the features. The RF features importance is shown in Table VIII.

Despite of the low R^2 result of LR, its ANOVA result of LR revealed p-value of 0.00, and no indication of multicollinearity, with VIF values ranged from 1.059 to 1.406. These indicate a good regression fit. Based on the coefficients of the LR model, the equation of *ppl* is as follows:

$$ppl = -0.1088 * pos + -0.041 * flg + 0.0015 * posd + 0.0004 * flgd + -0 * bl + -0.0176 * pic + -0.0019 * lin + -0.0019 * ni + -0.0013 * lt + 0.0013 * hc + -0 * pi + -0.0303 * hs + 0.085$$

V. Prediction Reliability

Prediction interval (PI) is a way to measure reliability for new instances. For tree-based regression models, such as Random Forest, quantile regression forest (QRF) is a way to generate prediction interval [33]. Compared to R^2 and error

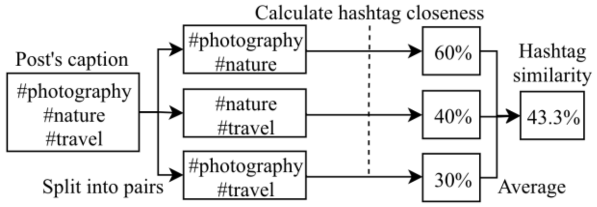


Fig. 2. Hashtag Similarity Score Example

measures, PI contains the dispersion of observations.

TABLE III
PERFORMANCE MEASURES

Indicator	Short	Formula	Value
R-squared	R^2	$\left(\frac{\sum_{i=1}^{n}(P_i - \bar{P})(A_i - \bar{A})}{n.S_P.S_A}\right)^2$	Closer to 1 is better
Mean absolute error	MAE	$\frac{\sum_{i=1}^{n}\lvert P_i - A_i\rvert}{n}$	Closer to 0 is better
Root mean square error	RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(P_i - A_i)^2}{n}}$	Closer to 0 is better
Relative absolute error	RAE	$\frac{\sum_{i=1}^{n}\lvert P_i - A_i\rvert}{\sum_{i=1}^{n}\lvert \bar{A} - A_i\rvert}$	Closer to 0 is better

Where:
n = Number of data, S = Std. deviation
P = Predicted value, A = Actual value

TABLE IV
DESCRIPTIVE STATISTICS OF FEATURES

Var	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
pos	0	1	0.04	0.08	4.75	31.47
flg	0	1	0.17	0.20	1.94	3.56
posd	-9.3	184.68	2.56	10.20	8.33	92.81
flgd	-99.86	1822.07	15.58	70.03	8.22	105.29
bl	0	507	56.14	62.85	0.99	0.51
pic	0	1	0.96	0.20	-4.63	19.45
lin	0	1	0.21	0.41	1.40	-0.05
cl	0	3765	105.31	173.22	3.98	31.41
ni	0	1	0.19	0.25	1.36	1.07
lt	0	1	0.24	0.31	1.10	-0.12
hc	0	30	0.45	0.95	8.09	143.27
pi	0	4096.67	455.85	651.82	2.39	6.62
hp	0	1	0.01	0.03	11.52	199.89
hs	0	1	0.01	0.01	30.14	1799.04
ppl	0	0.56	0.06	0.06	6.34	42.29

TABLE V
FEATURES AVERAGE FOR EACH USER TIER

Tier	*pos*	*flg*	*posd*	*flgd*	*bl*	*pic*	*lin*	*cl*	*ni*	*lt*	*hc*	*pi*	*hp*	*hs*
1	0.05	0.19	1.8	6	56	0.96	0.21	105	0.19	0.24	0.45	467	0.01	0.01
2	0.03	0.14	3.6	30	60	0.97	0.24	113	0.23	0.26	0.51	451	0.01	0.01
3	0.02	0.09	5.1	88	59	0.94	0.26	130	0.19	0.21	0.53	230	0.01	0.01
4	0.02	0.07	6.6	103	51	0.89	0.23	106	0.20	0.18	0.46	196	0.01	0.00
5	0.02	0.09	5.7	139	55	0.94	0.23	104	0.16	0.15	0.49	149	0.01	0.01
6	0.02	0.05	12.6	153	59	0.86	0.24	145	0.18	0.17	0.32	167	0.01	0.00
7	0.02	0.09	11.8	180	47	0.92	0.30	95	0.12	0.20	0.41	69	0.01	0.00
8	0.02	0.06	12.6	126	77	0.96	0.22	171	0.14	0.20	0.72	130	0.01	0.01
9	0.00	0.00	28.9	282	0	0.82	0.00	44	0.07	0.08	0.27	225	0.00	0.00

TABLE VI
USER TIERS CLASSIFICATION

Tier	1	2	3	4	5	6	7	8	9
Min ppl	0	0.062	0.124	0.186	0.248	0.31	0.371	0.433	0.495
Max ppl	0.062	0.124	0.186	0.248	0.31	0.371	0.433	0.495	0.557
Count Data	48,464	8,517	692	218	105	58	37	23	867

TABLE VII
PERFORMANCE COMPARISON

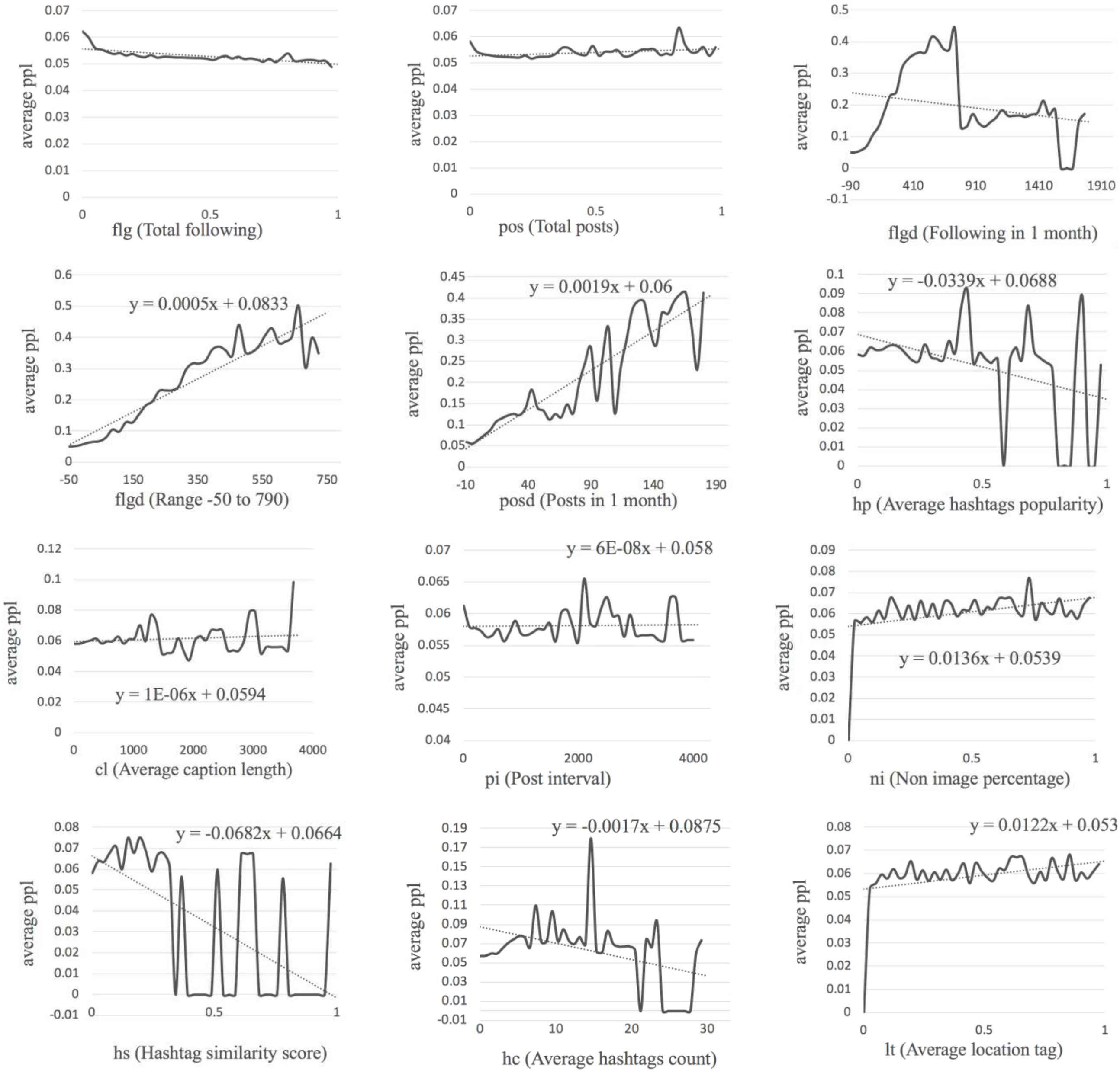
Indicator	LR	MARS	NN	XGBoost	RF	SVR
R^2	0.359	0.469	0.702	0.835	0.852	0.664
MAE	0.022	0.023	0.019	0.307	0.011	0.013
RMSE	0.048	0.044	0.034	0.310	0.023	0.035
RAE	0.971	0.994	0.819	13.404	0.464	0.553

TABLE VIII
FEATURE IMPORTANCE (FROM RANDOM FOREST)

Rank	Feature	Entropy (Avg. impurity decrease)	Nodes using the feature
1	*flg*	0.0490	237,760
2	*pos*	0.0166	95,320
3	*flgd*	0.0108	354,389
4	*posd*	0.0096	112,206
5	*pic*	0.0066	11,412
6	*hp*	0.0032	41,915
7	*bl*	0.0029	227,593
8	*pi*	0.0027	231,406
9	*cl*	0.0022	233,858
10	*lin*	0.0017	39,244
11	*ni*	0.0016	133,745
12	*hs*	0.0014	28,867
13	*hc*	0.0013	119,188
14	*lt*	0.0012	121,799

TABLE IX

PARTIAL DEPENDENCY PLOTS OF EACH FEATURE, ORDERED FROM MOST IMPORTANT FEATURE (Average of predicted *ppl* as the Y-Axis)



We generated the QRF using RF with the number of trees of 1,000 and 95% PI, using 10-folds cross validation. In QRF, for each data row, the result of each leaf (or estimator) will be recorded [33]. Each actual value will be checked whether it lies between 2.5% to 97.5% percentile of the leaves' results.

Based on this experiment, 90.1% of the actual values are within the PI. This indicates the percentage of reliability for future predictions. The error lines shown in Figure 3 visualize the interval of predicted values from every leaves, in which smaller interval indicates a more accurate result.

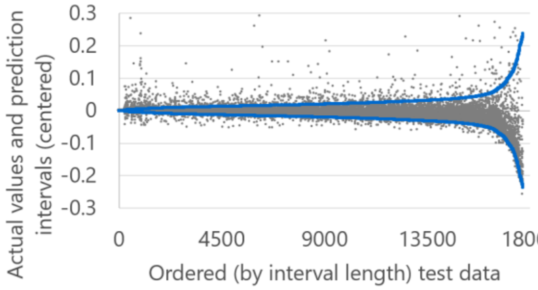


Fig. 3. Actual Values vs. Prediction Intervals generated with QRF. This was generated using 70% training set instead of cross validation, and there are 90.52% actual values (the scatter plots) fall within the interval.

VI. FEATURES IMPORTANCE ANALYSIS

From the regression results, there are two ways to analyze importance, i.e. using coefficients from LR and feature importance from RF. However, the LR model produced a low R^2 value. The feature importance from RF, on the other hand, doesn't show the negative or positive impact of each feature. To further analyze the features, a partial dependency plot (PDP) was used. The PDP can be used to examine each feature's partial contribution to the prediction result, taking into account the average effects of other predictors [34].

The PDP, as presented in Table IX, shows a feature as the X-axis, and the average of the predicted results as the Y-axis. The PDP results can be associated with RF's features importance (from Table VIII). In the PDP plot, the linear trend is added for each feature.

The {*flg*, *pos*} features show a very insignificant change, as opposed to{*flgd*, *posd*} which cause *ppl* to rise. This indicates that activeness is a key factor in raising popularity, as opposed to having a high number of posts and following in the past. A previous study also showed that bot-automated users can obtain +367% followers growth in 4 months [35]. Every +1 point of *posd* will increase *ppl* by +0.0019.

Based on the *flgd*, the *ppl* starts to decline at *flgd*=790. This indicates that following too many people in a month is spammy and has less value in raising popularity. However, when a trend line is drawn for *flgd* with a maximum range of 790, the trend line shows that every +1 point of *flgd* will increase *ppl* by +0.0005.

Adding a profile picture (*pic*) and a link (*lin*) will increase popularity. Users with profile picture have an average *ppl* of 0.084, compared to 0.061 if none. Users with a link have an average *ppl* of 0.073, compared to 0.064 if none.

The hashtag features, namely {*hp*, *hc*} cause a decrease of *ppl*. For *hp*, every +1 point will decrease *ppl* by -0.0146. This shows that using popular hashtags will not help in raising popularity. For *hc*, the plot shows a decline of *ppl* at some point. Additional analysis will be conducted.

Usage of video or carousel posts help in raising popularity. The trend line of *ni* shows that every +1 point of *ni* will increase *ppl* by +0.0136. The remaining features, namely {*hs*, *lt*} have a low importance in the prediction. Furthermore, the plot of *hs* highly varied, and the plot of *lt* showed an insignificant increase.

Overall, it can be concluded there are features that help in increasing popularity, and there are insignificant features.

TABLE X
FEATURES THAT AFFECT POPULARITY

Effect	Features and Description
Increases popularity	- Actively posting (*posd*) and following (*flgd*) - Adding profile picture (*pic*) and biography link (*lin*) - Using video and carousel posts (*ni*), instead of image - Using the right number of hashtag (*hc*) (11 is the best)
Decreases popularity	- Using popular hashtags (*hp*) decreases popularity.
Insignificant effect	- Using longer biography (*bl*) or longer captions (*cl*) has a very low impact - Number of posts (*pos*) and following (*flg*) in the past, or being inactive, don't affect popularity - Post interval (*pi*) doesn't affect popularity
Very low importance	Using similar hashtags (*hs*), using location tag (*lt*) affect popularity. However, due to the very low importance, the effects are negligible.

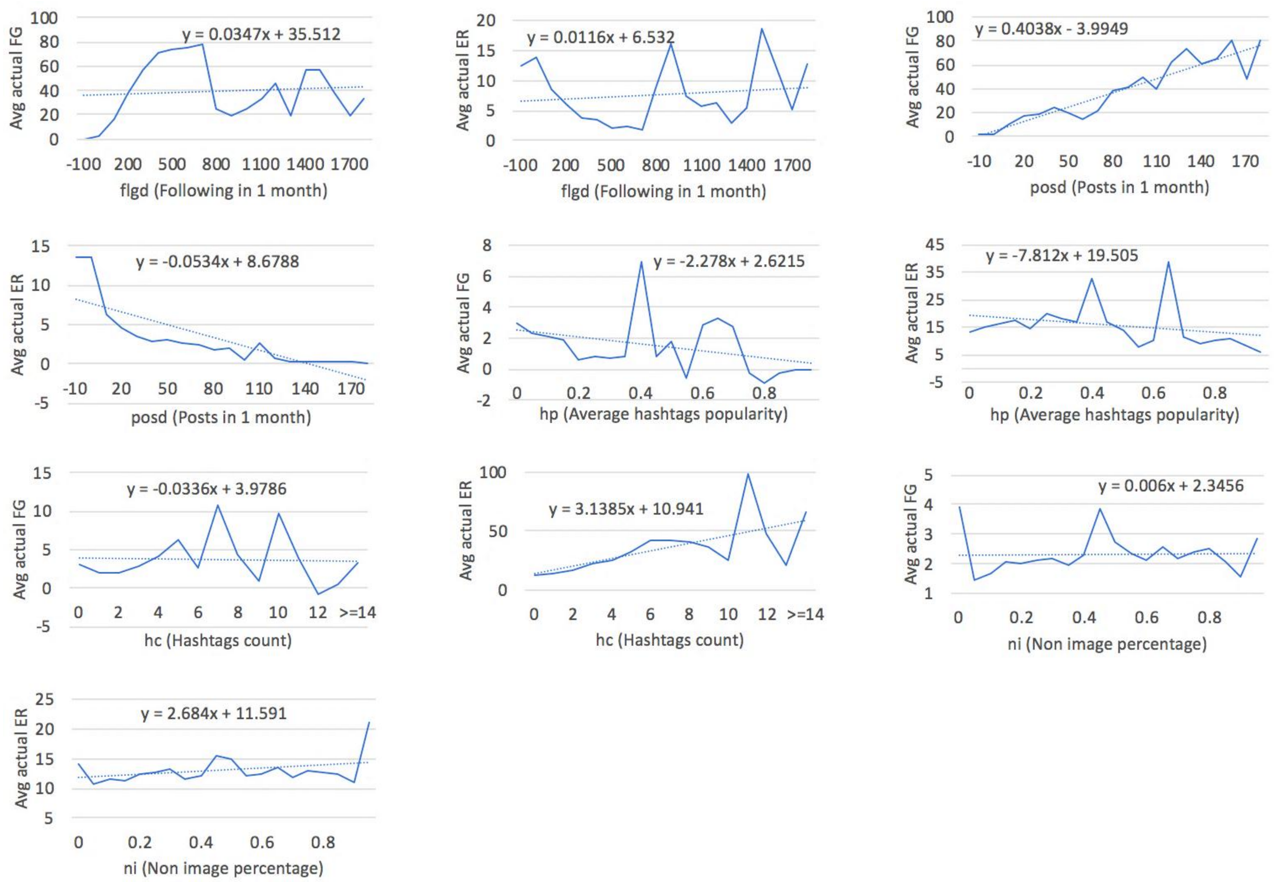
The conclusion of the effects is shown in Table X. The PDP uses predicted popularity as the Y-axis. The results can be compared to user tier analysis (Table V), which uses actual popularity as the measure. For example, the {*posd*, *flgd*} are indeed the most important predictors.

VII. FOLLOWERS GROWTH AND ENGAGEMENT RATE

The previous sections used *ppl* as the popularity measure, which might be difficult to interpret. This section is used to analyze the effect of the significant features, namely {*flgd*, *posd*, *hp*, *ni*, *hc*}, on followers growth (*FG*) and engagement rate (*ER*), and both are based on the actual (not predicted) values. All of the comparisons are shown in Table XI.

Actively following other people (*flgd*) doesn't affect ER, but causes the rise of FG, until *flgd* around 790. Consistent with the PDP result, this means that following too many

TABLE XI
COMPARISON OF SIGNIFICANT FEATURES VS. FG AND ER



people in a month has a bad effect on popularity. Actively posting (*posd*), on the other hand, constantly raises FG but decreases ER. This means that, while it increases the visibility of the user, the engagement becomes lower since the new followers are usually not close friends.

Using popular hashtags (*hp*) causes the decline of both ER and FG. Similar to PDP result, using popular hashtags has a bad effect since a post can be buried faster in the posts list and thus reduces visibility. Using more hashtags (*hc*), on the other hand, causes the rise of ER, but not FG. At *hc*=11, the average ER is 98.7%, which is the highest point. This means that the number of ideal hashtags for a post is 11. Surprisingly, this result is very similar to a report [36]. One last feature is the *ni*, which causes a little increase of FG and ER, similar to the PDP result.

VIII. Experiment on Features

This section is intended to discuss the possibility of predictions without certain features, especially those that are difficult to obtain. Only Random Forest model is used in this section, as the most accurate model according to the earlier experiment. The results are presented in Table XIII. Overall, it is shown that *hs* feature is the most insignificant and can be removed from the prediction model, while {*posd, flgd*} are the most significant.

A. Remove Features with High Correlations

Correlations analysis is commonly used to check multi-collinearity among the features [37]. When two features are highly correlated, one can be removed since using the other one is sufficient [38] [39].

Based on Pearson's correlation, multi-collinearity can be categorized to low correlation (for value < 0.4), medium correlation (for value between 0.4 and 0.7), and strong correlation (for value >= 0.7) [40]. As shown in Table XII, variables with the biggest correlation are *hp-hc* (0.45), *posd-pos* (0.411), *cl-bl* (0.410) lin-bl (0.371).

The features {*hp, posd, cl, lin*} were removed for this experiment. These features are harder to get if compared to the correlated partners; for example, *posd* requires two periods of scraping. As for *lin*, it has limited Boolean value. The result shows that the R^2 is down to 0.845.

B. Remove Hashtag Similarity (hs)

Hashtag similarity (*hs*) feature is the most time-consuming feature to calculate. Each two pair of hashtags in a caption was iterated through all hashtags (473k) in the dataset. Furthermore, it was shown in features importance that this was the third last important feature. Result shows that the decrease of R^2 is very insignificant, which shows that *hs* is not effective in the prediction.

C. Remove Features Requiring Second Period Scraping

The {*posd, flgd*} required re-scraping of users data in period 2. Due to their high importance ranks, it is shown in that the R^2 is down to 0.797, which is quite significant.

TABLE XII
CORRELATION ANALYSIS RESULT

	pos	*flg*	*posd*	*flgd*	*bl*	*pic*	*lin*	*cl*	*ni*	ltt	*hc*	*pi*	*hp*	*hs*
pos	1	0.2	0.4	0	0.3	0.1	0.2	0.2	0.1	0.2	0.1	-0.1	0.1	0.1
flg	0.2	1	0	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0	0	0	0
posd	0.4	0	1	0.3	0.1	0	0.1	0.1	0.1	0	0	-0.1	0	0
flgd	0	0.1	0.3	1	0	0	0	0	0	0	0	0	0	0
bl	0.3	0.2	0.1	0	1	0.2	0.4	0.4	0.2	0.2	0.2	-0.1	0.2	0.2
pic	0.1	0.1	0	0	0.2	1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
lin	0.2	0.1	0.1	0	0.4	0.1	1	0.3	0.1	0.1	0.1	-0.1	0.1	0.1
cl	0.2	0.1	0.1	0	0.4	0.1	0.3	1	0.2	0.2	0.2	-0.1	0.3	0.2
ni	0.1	0.1	0.1	0	0.2	0.1	0.1	0.2	1	0.3	0.1	0	0.1	0.1
ltt	0.2	0.1	0	0	0.2	0.1	0.1	0.2	0.3	1	0.2	0.1	0.2	0.2
hc	0.1	0	0	0	0.2	0.1	0.1	0.2	0.1	0.2	1	0	0.5	0.3
pi	-0.1	0	-0.1	0	-0.1	0.1	-0.1	-0.1	0	0.1	0	1	0	0
hp	0.1	0	0	0	0.2	0.1	0.1	0.3	0.1	0.2	0.5	0	1	0.3
hs	0.1	0	0	0	0.2	0.1	0.1	0.2	0.1	0.2	0.3	0	0.3	1

TABLE XIII
EXPERIMENT ON FEATURES - PERFORMANCE COMPARISON

Indicator	Exp. A	Exp. B	Exp. C	Exp. D
R^2	0.845	0.852	0.797	0.842
MAE	0.011	0.010	0.012	0.011
RMSE	0.024	0.022	0.026	0.023
RAE	0.474	0.464	0.560	0.480

D. Remove Media Data

The *features* {*cl, ni, lt, hc, pi, hp, hs*} are coming from media data. Removing them, while it decreases the R^2 to 0.842, it still produced a reasonable prediction model.

IX. Discussion

Instagram is the most popular platform for brand marketing. In this regard, the user's popularity becomes very important. Users with high engagement and a high number of followers become new influencers. This research can help business users to predict an influencer's popularity for marketing purpose. It also helps ordinary users to understand popularity factors.

This research has successfully created regression models for predicting a user's popularity in terms of followers growth (FG) and engagement rate (ER). The best model to predict popularity was Random Forest, with an accuracy of 85.2%, measured with R^2. This level of accuracy was able to deliver pre-analysis results (descriptive statistics) that are consistent with the post-analysis results (features importance, FG and ER, experiment on features).

In both pre-analysis and post-analysis, it was shown that there were features that can significantly increase popularity, especially being active in posting and following users (*posd* and *flgd* features). Users have to be active, instead of relying on their existing posts and following, even though the numbers are high. Other significant features were completing metadata, using video or carousel posts. and using 11 hashtags in a post.

To our surprise, using popular hashtags (*hp* feature) does not help in increasing popularity, both in terms of followers growth and engagement rate. This shows that users need to increase the post's quality, instead of using hashtags trick. Furthermore, due to the large volume of posts in a popular hashtag, a new post can be buried faster in the posts list.

There were also features which have an insignificant effect, i.e. number of total posts (*pos*) and following (*flg*), post interval (*pi*), and features with low importance, i.e. usage of similar hashtags (*hs*), and location tag (*lt*).

Furthermore, the experiment section showed that *hs* could be removed.

X. Conclusion

This research used metadata, media, hashtag popularity and similarity as the features for prediction. The hashtag analysis, as well as user's popularity prediction (as opposed to post's popularity), are still non-existent in recent studies. With the prediction accuracy of 85.2% and reliability of 90.1% (using 95% prediction interval), the produced Random Forest model will be accurate enough for practical use.

For future work, methods to predict the authenticity or emotion of users can be incorporated, such as sentiment analysis, fake accounts detection [41], and malicious content detection [42]. It was proven that non-authentic users can behave differently from authentic users [41]. Image analysis can also be added, such as the image quality and category of a post. There were studies which suggested that the category of pictures is highly related to the number of likes or followers [43] [44].

Existing studies still lacked the biography text analysis, such as the attractiveness of the user's biography. This research showed that the user's metadata has a significant effect on popularity. However, the biography length factor only sits as the 7th most important feature in the prediction. Thus, we believe that additional text analysis to the biography will raise its importance as other metadata do.

References

[1] B. E. Weeks, A. Ardèvol-Abreu and H. G. d. Zúñiga, "Online influence? Social media use, opinion leadership, and political persuasion," *International Journal of Public Opinion Research,* vol. 29, no. 2, pp. 214-239, 2017.

[2] O. E. Ogunyombo, O. Oyero and K. Azeez, "Influence of Social Media Advertisements on Purchase Decisions of Undergraduates in Three Nigerian Universities," *Journal of Communication and Media Research,* vol. 9, pp. 244-255, 2017.

[3] M. Moussaïd, J. E. Kämmer, P. P. Analytis and H. Neth, "Social Influence and the Collective Dynamics of Opinion Formation," *PloS one,* vol. 8, no. 11, p. https://doi.org/10.1371/journal.pone.0078433, 2013.

[4] Statista, "Most famous social network sites worldwide as of July 2018, ranked by number of active users (in millions)," 2018. [Online]. Available: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. [Accessed 3 October 2018].

[5] D. Tanase, D. Garcia, A. Garas and F. Schweitzer, "Emotions and Activity Profiles of Influential Users in Product Reviews Communities," *Frontiers in Physics,* vol. 3, no. 87, p. doi: 10.3389/fphy.2015.00087, 2015.

[6] S. Khamis, L. Ang and R. Welling, "Self-branding, 'micro-celebrity' and the rise of Social Media Influencers," *Celebrity Studies,* vol. 8, no. 2, pp. 191-208. https://doi.org/10.1080/19392397.2016.1218292, 2017.

[7] A. Clasen, "Instagram 2015 Study – Unleash the Power of Instagram," 2015. [Online]. Available: https://blog.iconosquare.com/instagram-2015-study-unleash-power-instagram/. [Accessed 5 November 2018].

[8] E. Akar, H. F. Yüksel and Z. A. Bulut, "The Impact of Social Influence on the Decision-Making Process of Sports Consumers on Facebook," *Journal of Internet Applications & Management,* vol. 6, no. 2, 2015.

[9] I. Roelens, P. Baeckeb and D. Benoita, "Identifying influencers in a social network: The value of real referral data," *Decision Support Systems,* vol. 91, pp. 25-36. http://dx.doi.org/10.1016/j.dss.2016.07.005, 2016.

[10] S. Barker, "What's the Difference Between Celebrities and Influencers – and Which Does Your Brand Need?," 2018. [Online]. Available: https://smallbiztrends.com/2018/02/influencers-vs-celebrities.html. [Accessed 5 November 2018].

[11] S. Utz, M. Tanis and I. Vermeulen, "It Is All About Being Popular: The Effects of Need for Popularity on Social Network Site Use," *Cyberpsychology, Behavior, and Social Networking,* vol. 15, no. 1, pp. 37-42, 2012.

[12] J. Ehrhardt, "Sussing Out Growth - How to Interpret Follower Growth Rates," 17 May 2017. [Online]. Available: https://blog.influencerdb.com/follower-growth-rates/. [Accessed 9 July 2019].

[13] I. CO, "How Long Does It Take To Become An Influencer? Meet The Instagrammer Who Became a Brand Ambassador In 2 Months," 2017. [Online]. Available: https://influencercollective.com/how-a-micro-influencer-on-instagram-became-an-ambassador-for-becca-scrub/. [Accessed 9 July 2019].

[14] University of Waikato, "WEKA: The workbench for machine learning," [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. [Accessed 10 March 2020].

[15] G. D. Saxton, J. N. Niyirora, C. Guo and R. D. Waters, "#AdvocatingForChange: The Strategic Use of Hashtags in Social Media Advocacy," *Advances in Social Work,* vol. 16, no. 1, pp. 154-169, 2015.

[16] K. Cakmak, I. Cikrikcioglu, O. Demiralp, A. Ozturk, F. S. Palut, Y. Yilancioglu and M. Yildirim, "The Causal Determinants of Popularity in Instagram," 2017.

[17] E. G. Martın, N. Lavesson and M. Doroud, "Hashtags and followers : An experimental study of the online social network Twitter," *Journal of Marketing Management,* vol. 31, no. 1, pp. 221-243, 2013.

[18] C. J. Qian, J. D. Tang, M. A. Penza and C. M. Ferri, "Instagram Popularity Prediction via Neural Networks and Regression Analysis," in *IEEE Transactions on Multimedia 19.11,* 2017.

[19] S. Ayres, "Do Instagram Hashtags Really Lead to More Engagement?," 2017. [Online]. Available: https://www.agorapulse.com/social-media-lab/instagram-hashtags-engagement. [Accessed 12 March 2019].

[20] K. Burney, "Everything Marketers Need To Know About Instagram Sponsored Content," 2015. [Online]. Available: https://trackmaven.com/blog/everything-marketers-need-know-instagram-sponsored-content/. [Accessed 12 March 2019].

[21] N. L. Khalid, S. Y. Jayasainan and N. Hassim, "Social media influencers - shaping consumption culture among Malaysian youth," in *SHS Web of Conferences Vol. 53,* 2018.

[22] P. C. Austin, "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality," *Statistics in Medicine,* vol. 26, p. 2937–2957, 2007.

[23] W. Zhang and A. T. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile

drivability," *Geoscience Frontiers,* pp. 1-8, 2014.

[24] C. Acciani, V. Fucilli and R. Sardaro, "Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach," *Aestimum,* pp. 27-45, 2011.

[25] N. S. Hussien, S. Sulaiman and S. M. Shamsuddin, "Tools in data science for better processing," in *AIP Conference Proceedings 1750, 020017,* 2016.

[26] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining,* ACM, 2016, p. 785–794.

[27] "The Fake project," [Online]. Available: http://wafi.iit.cnr.it/theFakeProject/ . [Accessed 30 October 2015].

[28] M. Marzjarani, "Sample Size and Outliers, Leverage, and Influential Points, and Cooks Distance Formula," *International Journal of Arts and Commerce,* vol. 4, no. 2, pp. 83-86, 2015.

[29] C. Laurence, "How Do I Calculate My Engagement Rate on Instagram?," Plann, 2019. [Online]. Available: https://www.plannthat.com/calculate-engagement-rate-on-instagram/. [Accessed 2 November 2019].

[30] J. Clement, "Worldwide Instagram follower growth rate from January to June 2019, by profile size," Statista, 1 October 2019. [Online]. Available: https://www.statista.com/statistics/307026/growth-of-instagram-usage-worldwide/. [Accessed 2 November 2019].

[31] A. Z. Ul-Saufie, A. S. Yahya, N. A. Ramli and H. A. Hamid, "Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM 10 Concentration Level Based On Gaseous And Meteorological Parameters," *International Journal of Applied Science and Technology,* vol. 1, no. 4, pp. 42-49, 2011.

[32] E. Tosun, K. Aydin and M. Bilgili, "Comparison of linear regression and artificial neural network model of a diesel engine fueled with biodiesel-alcohol mixtures," *Alexandria Engineering Journal,* vol. 55, p. 3081–3089, 2016.

[33] N. Meinshausen, "Quantile Regression Forests," *Journal of Machine Learning Research,* vol. 7, 2006.

[34] B. M. Greenwell, "pdp: An R Package for Constructing Partial Dependence Plots," *The R Journa,* vol. 9, no. 1, pp. 421-436, 2017.

[35] P. Bellavista, L. Foschini and N. Ghiselli, "Analysis of Growth Strategies in Social Media: the Instagram Use Case," 2019.

[36] L. Myers, "How to Use Hashtags on Instagram for Amazing Growth in 2020," Louise Myers Visual Social Media, 3 February 2020. [Online]. Available: https://louisem.com/7198/how-to-use-hashtags-on-instagram. [Accessed 11 March 2020].

[37] T. Napierala and A. Pawlicz, "The determinants of hotel room rates: an analysis of the hotel industry in Warsaw, Poland," *International Journal of Contemporary Hospitality Management,* vol. 29, no. 1, pp. 571-588, 2017.

[38] D. C. Montgomery, E. A. Peck and G. G. Vining, Introduction to Linear Regression Analysis, John Wiley & Sons, Hoboken, New Jersey, United States: John Wiley & Sons, Inc., 2012.

[39] A. Rebekić, Z. Lončarić, S. Petrović and S. Marić, "Pearson's or spearman's correlation coefficient – which one to use?," *Poljoprivreda (Osijek),* vol. 21, no. 2, pp. 47-54, 2015.

[40] D. S. Moore, W. I. Notz and M. A. Fligner, The basic practice of statistics, New York: W. H. Freeman and Company, 2013.

[41] K. R. Purba, D. Asirvatham and R. K. Murugesan, "Classification of instagram fake users using supervised machine learning algorithms," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 10, no. 3, pp. 2763-2772, 2020.

[42] P. Wanda and H. J. Jie, "DeepSentiment: Finding Malicious Sentiment in Online Social Network based on Dynamic Deep Learning," *IAENG International Journal of Computer Science,* vol. 46, no. 4, pp. 616-627, 2019.

[43] Y. Hu, L. Manikonda and S. Kambhampati, "What We Instagram: A First Analysis of Instagram Photo Content and User Types," in *Eighth International AAAI conference on weblogs and social media*, 2014.

[44] M. Mazloom, R. Rietveld, S. Rudinac, M. Worring and W. v. Dolen, "Multimodal popularity prediction of brand-related social media posts," in *Proceedings of the 24th ACM international conference on Multimedia*, ACM, 2016, pp. 197-201.

[45] Z. Zhang, T. Chen, Z. Zhou, J. Li and J. Luo, "How to Become Instagram Famous: Post Popularity Prediction with Dual-Attention," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018.

[46] I. L. d. Silva, "Hashtag popularity prediction for social networks," FEUP - Faculdade de Engenharia da Universidade do Porto, Portugal, 2018.

[47] S. De, A. Maity, V. Goel, S. Shitole and A. Bhattacharya, "Predicting the Popularity of Instagram Posts for a Lifestyle Magazine Using Deep Learning," in *2nd International Conference on Communication Systems, Computing and IT Applications (CSCITA)*, 2017.

[48] T. Trzcinski and P. Rokita, "Predicting popularity of online videos using Support Vector Regression," *IEEE Transactions on Multimedia,* vol. 19, no. 11, pp. 2561-2570, 2017.