What Features of a Twitter User are Linked to Popularity?

Ethan Husband, Ishaan Srivastava, Parsa Babadi Noroozi, James Garnham

Contents

Aim

The aim of this project is to employ supervised learning techniques in order to investigate which features of a twitter user are most significantly predictive of their follower base. This analysis has an obvious application for Twitter users (influencers, businesses, politicians etc.) and social media analysts, who may find these results beneficial for optimising growth and impressions.

Dataset

The dataset used in this report was acquired from another report investigating the difference between expert and non-expert twitter users[1], and contains mostly normalised numeric data on a sample of 5280 twitter users. There are a total of 70 features present in that dataset, of which 35 are analysed in this investigation and split into 4 categories. These categories include basic user information (e.g. number years on the app), language (e.g. number of periods and colons used per tweet), interactivity (e.g. messages with mentions and hashtags) and coherence (topic coherence, lexical coherence).

Preprocessing and Wrangling

As most features in the dataset took a small or normalised value between 0 and 1, many of the analysis techniques yielded highly skewed results due to a scale disproportionate to followers. Therefore, the decision was made to use the logarithmic (base 10) scale for the number of followers. Following this, it was also decided to impute any missing values using the mean for each field. Once the data was preprocessed, it was partitioned into the relevant categories and sent through an abstract processing pipeline[2] which performed analyses on all categories.

Analysis Methods

Methods

The entirety of the data relevant to our analysis was numerical. Consequently, the analysis techniques employed in this study were Pearson Correlation, Normalised Mutual Information, as well as Multilinear Regression. Consequently, the use of categorical analysis techniques was avoided with the exception of the NMI binning, since greater contextual knowledge of the dataset would be required to meaningfully group data. A brief contextual overview of the techniques used in the processing pipeline is as follows:

Pearson Correlation

To gauge the extent of a linear relationship between different factors and the number of followers, the Pearson correlation (ρ) was computed for each variable against a user's

number of followers. This aims to identify any linear trends/relationships. Note this correlation coefficient is *not robust* as it can be affected by outliers, however this is largely mitigated by taking the logarithm of followers.

Normal Mutual Information

To address the limitations of Pearson correlation only identifying linear relationships, Normalised Mutual Information (NMI) was calculated for all fields against the followers field. This measures the general reduction in uncertainty regarding what range a user's number of followers lies in, given a known value of the other variable. This typically applies to categorical datasets, so the relevant fields were discretized by binning each value into 10 different ranges.

Train-test-split

In order to employ supervised learning via regression analysis, the dataset must be split into a training set and a testing set. In this case, the data was split 4:1 - ⅘ of the data was used to train the model, and ⅕ used to test it.

Multivariate Regression

Finally, regression analysis was undertaken on each section of the dataset, in order to derive a linear model between the variables of the section and followers. This can then be used with the test to compare the predictions of the model against the actual values possessed by users - ultimately determining which sections were most accurately predictive of followers. Evaluation metrics for the model were mean squared error (MSE) between predictions and actual values, as well as an R^2 score, which measures the total variance described by the model in proportion to the actual variance of the dataset.

Discussion

1. Basic Features

Firstly, we intend to analyse basic features of the dataset. As per table 1.1, one can immediately see the feature which has the largest magnitude correlation is 'years on app' (ρ = 0.49), indicating a relatively strong positive linear relationship with followers. 'Number of friends' also has a proportionally high positive correlation of 0.22.

Meanwhile, the most negative correlation observed is -0.17 for the 'tweets per retweet' variable, indicating a potential negative linear relationship between 'ratio of tweets to retweets' and number of followers.

	followers
total_tweets	-0.019039
per_rt	-0.170546
friends	0.216338
years	0.493978

Table 1.1 Correlation Matrix for **Basic Features**

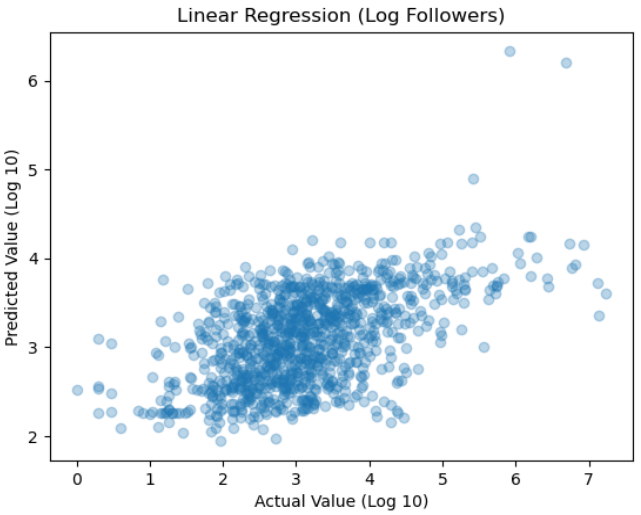
Interestingly, ‘total number of tweets’ (ρ = 0.02) had the weakest correlation observed with followers, a surprising value implying virtually no association between a user's number of followers and their number of tweets.

Next, we see the normalised mutual information values for basic features with followers**.** In the above table we see that “friends” (number of friends) has a relatively high NMI of 0.24, revealing that of all the basic features, knowledge in how many friends a user has most reduces the uncertainty in how many followers they have. The variable with the lowest NMI value was “total tweets” (which had the lowest magnitude correlation as well), again somewhat against intuition.

	Feature	NMI with followers
0	total_tweets	0.029514
1	per_rt	0.032828
2	years	0.094401
3	friends	0.235748

Table 1.2 NMI Matrix for **Basic Features**

Finally, we look at the regression output as well as the measures of fit. The clear linear trend in figure 1.3 bodes well for this model, as increases in the actual number of user followers tended to result in higher predictions, which would only occur if the features used for this model did share some linear relationship with followers. The R^2 value of 0.284 for this model, as well as it having the lowest MSE of all models, indicates that the linear model can be a somewhat accurate predictor. Of course, the residuals plot in Appendix A.1 does remind us that there are clear differences between the model’s predictions and the true data, and it is far from perfect.



Figures 1.3 Regression Analysis for **Basic Features**

2. Language

The dataset was further analysed according to ‘language’ features, comprising syntactic and linguistic characteristics including: frequency of various punctuation marks (commas, periods, etc.), number of characters, and overall punctuation frequency, both per tweet and per retweet (where ‘_rt’ signifies ‘per retweet’).

From the Pearson correlation matrix (table 2.1), it can be observed that correlation of various ‘language’ features with follower count ranged from negligible (<0.09) to moderately low (0.09-0.22). Notably, ’language’ features of *retweets* (semi_rt, periods_rt, colon_rt, etc.) produced consistently smaller absolute Pearson correlation values than those of *tweets themselves* (exmark, semi, commas, etc.), suggesting that punctuation frequency and character count in tweets were a better indicator of follower count than the same features in retweets.

	followers
punc_rt	0.003285
chars_rt	0.015662
colon_rt	-0.019071
quesmark	-0.020595
commas_rt	0.030114
periods_rt	-0.038660
quesmark_rt	0.048307
semi_rt	0.075530
exmark_rt	0.082600
exmark	0.091870
semi	0.092039
commas	0.097847
periods	0.124068
colon	0.168808
punc	0.177964
chars	0.213087

Table 2.1 Correlation Matrix for **Language**

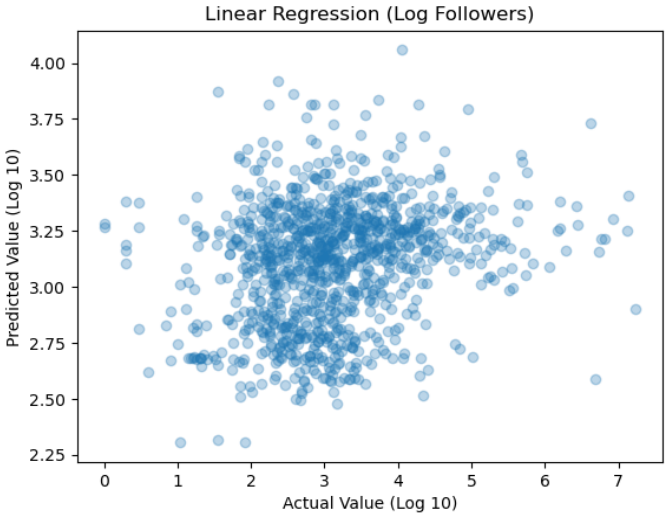
Of these ‘language’ features, the strongest correlations with follower base were both positive correlations: overall punctuation frequency in tweets (punc, ρ = 0.178) and number of characters per tweet (chars, ρ = 0.213). This suggests that more punctuation per tweet may indicate a greater number of followers, while longer tweets may also be indicative of a larger follower base, though the correlations are not strong.

The NMI matrix (table 2.2) produced no values >0.04, indicating that there was no major non-linear correlation between ‘language’ features that may have been overlooked by Pearson correlation analysis.

The relatively random distribution of the residual plot (figure A.2), suggests that a linear assumption is supported for this grouping of features. However, the regression’s R^2 score was low (0.0483) with a very high mean squared error of 1.03. Coupled with the lack of visible correlation between real and predicted values (figure 2.3), as well as the consistently large residual values (figure A.2), this suggests that the model could not predict a user’s followers based on these aspects of their language use with a high degree of accuracy, and that the analysed ‘language’ features together were ultimately a poor predictor of a user’s Twitter follower count.

	Feature	NMI with followers
0	punc_rt	0.015222
1	commas_rt	0.015302
2	periods_rt	0.015673
3	exmark_rt	0.016962
4	chars_rt	0.018820
5	quesmark_rt	0.023170
6	semi_rt	0.023417
7	chars	0.024671
8	semi	0.026622
9	exmark	0.029757
10	colon_rt	0.030169
11	punc	0.030281
12	colon	0.031298
13	commas	0.033701
14	periods	0.038153
15	quesmark	0.038224

Table 2.2 NMI Matrix for **Language**



Figures 2.3 Regression Analysis for **Language**

3. Interactivity

	followers
percent_msgwithtag	-0.028242
percent_msgwithment_rt	0.030090
urlpermsg_rt	0.035741
percent_msgwithtag_rt	0.053940
percent_msgwithurl_rt	0.064057
tagpermsg	-0.110292
mentpermsg	0.114338
percent_msgwithment	0.147174
percent_msgwithurl	0.169069
mentpermsg_rt	0.170499
tagpermsg_rt	-0.180448
urlpermsg	0.189626
followers	1.000000

	Feature	NMI with followers
0	percent_msgwithment_rt	0.017630
1	percent_msgwithurl_rt	0.019739
2	percent_msgwithtag_rt	0.021342
3	urlpermsg_rt	0.023954
4	percent_msgwithtag	0.028459
5	percent_msgwithment	0.029148
6	mentpermsg	0.031460
7	tagpermsg	0.031946
8	percent_msgwithurl	0.032483
9	mentpermsg_rt	0.036509
10	urlpermsg	0.042057
11	tagpermsg_rt	0.052765

Table 3.1 Correlation Matrix for **Interactivity** **Table 3.2** NMI Matrix for **Interactivity**

From the correlation matrix (table 3.1) it is clear that the top seven features have correlations ranging from $|0.11|$ to $|0.19|$ with a big gap between the seventh and eighth features. The matrix potentially highlights the significance of urls in original messages along with tags and mentions in retweets.

The correlation matrix (table 3.2) also indicates that out of the 12 total features in the interactivity category only three are negatively correlated, suggesting that interactivity features tend to move in the same direction as the number of followers.

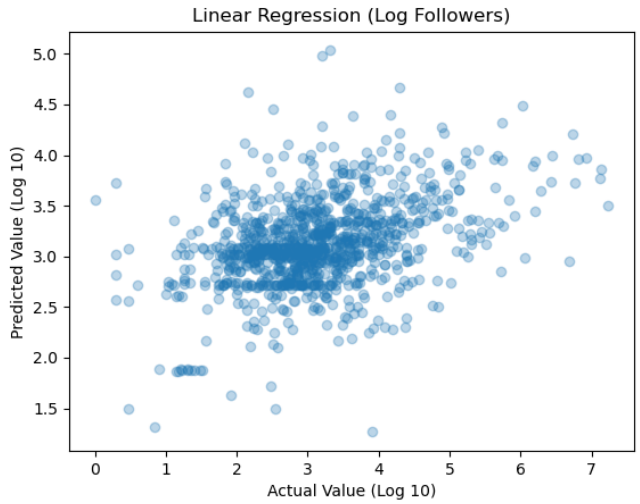
The NMI matrix depicts the information gained about followers from other features in the interactivity category. The mutual information of the features range from 0.018 to 0.053, which could suggest that there is no major information gain from any of the information given that the maximum value for NMI is 1. However the order of the top 4 features in the NMI matrix is almost identical to that of the correlation matrix with only the first and second swapping places. This could potentially reinforce the relationships between the number of followers and these features as also seen in the correlation matrix.

From the linear regression plot (figure 3.3), there only seems to be very little linear relationship between the actual and predicted values.

The residuals plot (figure A.3) seems to have a random distribution and is mostly symmetrical. Beside the central cluster where most of our residuals reside, there is no clear pattern in the graph which could potentially indicate that a linear model such as the one being used is appropriate for modelling the data.

However from previous analysis and our predicted vs actual values, the high mean squared error of 0.92 along with a low coefficient of determination of 0.15 we can conclude that the

model developed is not a very good fit for modelling the relationship between interactivity features and number of followers.



Figures 3.3 Regression Analysis for **Interactivity**

4. Coherence

Here we analyse the dataset on the basis of 'coherence', and its statistical relevance to the corresponding number of user followers. To achieve this, we isolate the fields of lexical coherence (or lexco) and topic coherence from the dataset, both of which were derived in the study[1] the dataset originated from and take a value from 0 to 1.

At first glance, the correlation matrix (table 4.1) shows a small negative correlation for both features. Although, within the context of correlation results in other sections, the result for lexical coherence is quite significant, being one of the strongest negative correlations and highest NMI values (table 4.2) thus far. The comical nature of this is entirely acknowledged.

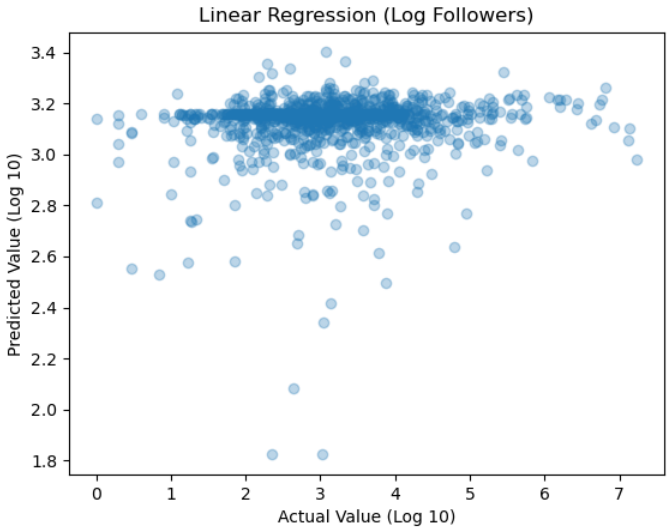
	followers
topic_coherence	-0.065620
lexco	-0.120001

Table 4.1 Correlation Matrix for **Coherence**

	Feature	NMI with followers
0	topic_coherence	0.020487
1	lexco	0.062927

Table 4.2 NMI Matrix for **Coherence**

Performing regression analysis with only coherence features results in a poorer model, almost always predicting around 10^3 - $10^{3.2}$ followers. This is likely due to topic coherence having many imputed values, so naturally the algorithm would predict a similar value for all those entries with identical imputations.
Funnily, one can also notice in figure 4.3 a small number of outlier users which were predicted to have < 10^2 followers. A closer look into these data entries would reveal that these users had the maximal lexical coherence.



Intercept:
3.24789
Coefficients:
[Topic, Lexco]:
[-1.33470 -1.67739]
Regression R score:
0.00707
Regression Mean Squared Error (MSE):
1.07467

Figures 4.3 Regression Analysis for **Coherence**

Ultimately, coherence acts as a weak predictor for followers, with the regression analysis having an R^2 score of roughly 0.001 and the highest MSE yet, which is extremely poor. However, one must consider whether this analysis may be undermined by the poor data integrity of topic coherence. Especially given the relative strength of the negative correlation possessed by lexical coherence, further investigation ought to be taken into the significance of lexical coherence on its own.

Conclusion

While many features demonstrated little correlation with a user's number of followers, ultimately, basic, 'identity-based' features, principally 'years since joining Twitter', were found to be the strongest indicators of follower count, relative to other features based on language, interactivity and coherence. The primary conclusion for interested parties is that Twitter user popularity is less associated with the nature of tweets and retweets, but more so related to time spent growing a following.

Evaluation

While overall, the investigation did help identify critical variables relevant to a follower base, the limitations of the analysis techniques must be acknowledged.
Firstly, correlation and regression analysis only act to identify linear relationships, which may be problematic as our investigation is oriented towards identifying *any* relationship between a feature and a user's follower base, not just linear relationships. Using mutual information mitigates this issue somewhat, but this does not specify the nature of the relationship, just the existence of one.

Secondly, a further limitation is related to the binning associated with NMI calculation. The same binning technique is applied to all data fields by splitting them into 10 bins of equal

width. This is not contextually appropriate. To bin the data fields correctly one should examine each data field individually and apply contextual knowledge pertinent to the field.

Finally, another issue arose through imputation of the mean for missing values. While mean imputation enables us to use entries with missing field values, in turn, this can skew the variance of the data to be smaller than it would be in reality. In particular, this can significantly impact the accuracy of the regression analysis, as the model becomes more likely to predict the mean value which was precisely seen in the analysis of coherence.

Appendix

Appendix A: Residual Plots for regression analysis

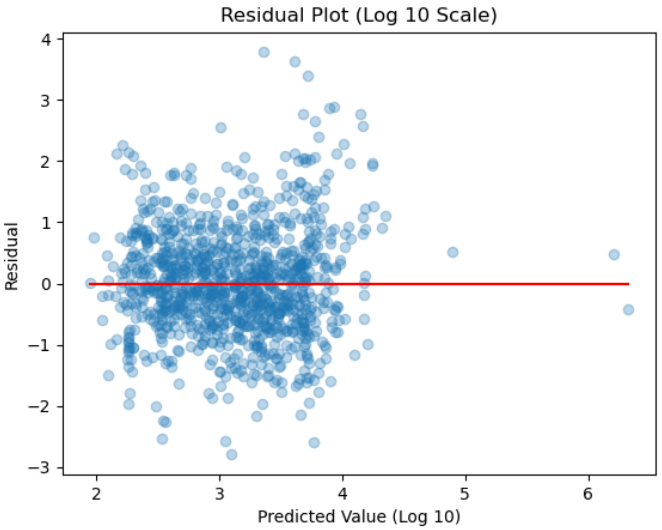


Figure A.1 Residual Plot for Basic Features

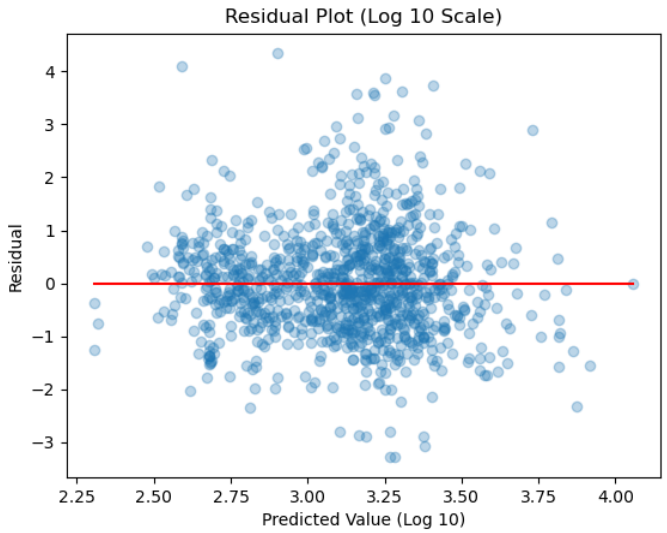


Figure A.2 Residual Plot for Language Features

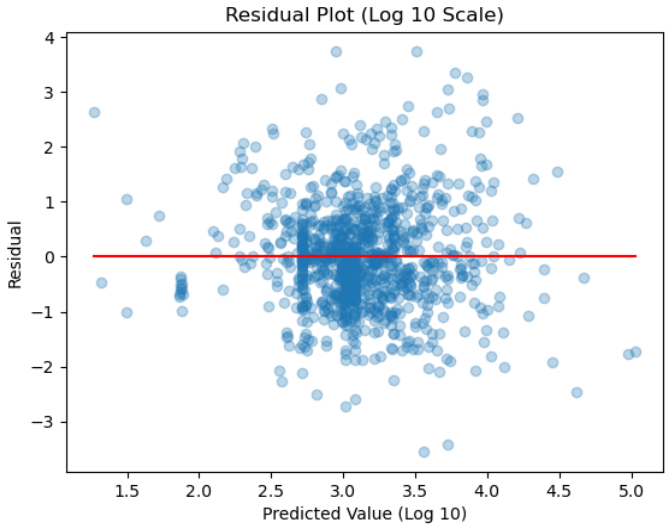


Figure A.3 Residual Plot for Interactivity Features

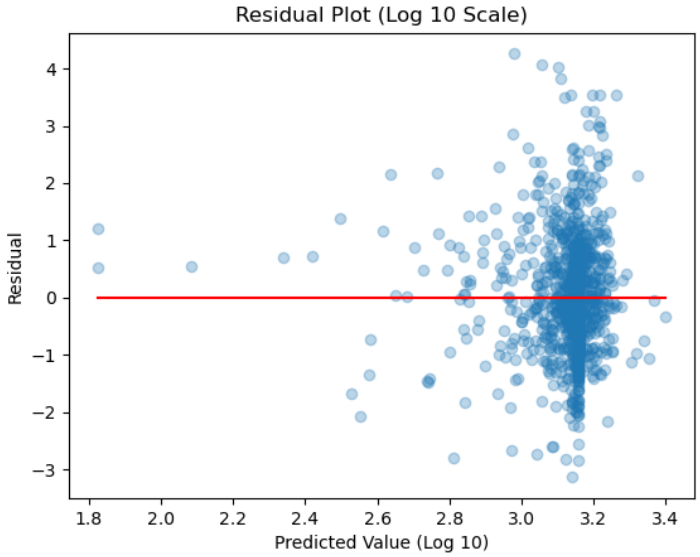


Figure A.4 Residual Plot for Coherence Features

References

[1] Horne, B. D., Nevo, D., Freitas, J., Ji, H., & Adali, S. (2016). *Expertise in social networks: How do experts differ from other users?* Retrieved October 16, 2022, from https://blender.cs.illinois.edu/paper/expertise2016.pdf

[2] Husband, E., Srivastava, I., Babadi Noroozi, P., Garnham, J. (2022). What Features of a Twitter User are Linked to Popularity? https://github.com/ethanhusband/follower-analysis