Machine Learning Approach to Analyze and Predict the Popularity of Tweets with Images

Nimish Joseph, Amir Sultan, Arpan Kumar Kar, P. Vigneswara Ilavarasan

Machine learning approach to analyze and predict the popularity of tweets with images

Nimish Joseph[1][0000-0002-0560-7325], Amir Sultan[0000-0002-9928-8037], Arpan Kumar Kar[0000-0003-4186-4887] and P Vigneswara Ilavarasan[0000-0002-9431-3520]

Department of Management Studies, IIT Delhi, India
[1]nimishjoseph@gmail.com

Abstract. Social Media platforms play a major role in spreading information. Twitter, is one such platform which is used by millions of people to share information every day. Twitter with the recent introduction of a feature that helps its users to attach images to a tweet has changed the dynamics of tweeting. Many people now prefer to tweet with images. This study tries to analyse and predict the popularity of such tweets. This study uses learning mechanisms like decision tree, neural networks and random forests to learn the tweets posted by people with a higher number of followers. Image parameters, network variables, transactional, and historical variables of a tweet are identified and are trained for predicting the test data. This study can help businesses to build better social media tools, which allows customers to tweet data at the right time. This study also identifies the contribution of various parameters that may help a tweet to go viral.

Keywords: Twitter image, neural network, random forest, decision tree, popularity, machine learning

1 Introduction

Millions of images are uploaded on internet every day and the number keeps on increasing with popularity of social media platforms [1]. Social media platforms like Snapchat, Instagram, Twitter and Facebook are basically used by its users to upload pictures. Various studies show that more than 50% users login to photo sharing platforms like Instagram and Snapchat daily [2]. Majority of the pictures uploaded are ignored by the large share of users. However, certain posts or activities with images gain the attraction of audience [3,4]. Twitter is one of the most widely used platforms by businesses. Twitter allows its users to share images along with the textual content. This feature of tweets has brought in more users and audience. Further, research shows tweets with media content have much better impressions, engagement, and shareability [5]. Hence, businesses try to incorporate images when they post content in twitter.

The increase in image content usage has brought the attention of academicians, researchers and businesses. Many tweets with image content performs better compared to those without images [6]. However, this is not always true. Various descriptive and content related studies of twitter data [7] explains about popularity with respect to textual data. These studies extract the text related variables in twitter data and perform the analysis. It is therefore necessary to understand that, with the change in tweeting pattern, do we have a change in parameters that will influence the popularity of tweets?

Businesses and individuals try to attach images to their tweets and with the better connectivity, users prefer to get clarity, viewing the image related to textual data [8]. The increase in number of views, viewers retweets, favorites and other activities will enhance the chances for a tweet to go popular. However, popularity also depends on the user profile [9]. A celebrity with a million followers posting a tweet and getting 100 favorites, may not be counted as a popular tweet. Instead a person with 200 followers getting a100 favorites for a tweet could be counted as popular. Hence, it is necessary to understand the history of users' tweets and its reach to define popularity of a tweet. Machine learning techniques are one of the best ways to perform classification [10] and the tweets in our study needs to be classified to identify the popularity.

This study, tries to identify the features of tweets with images, which could make it popular. This study will also establish a quantitative way to predict popularity of post with images on social media platforms like twitter just before uploading it based on transactional variables, image features, network variables, and historical social activities with the help of various machine learning models.

2 Literature Review

Many of the images that appear in SM goes viral [11]. This has really catapulted the popularity of image focused platforms like Pinterest and Instagram in recent times. For a category of images, studies tries to identify if there is a difference in the various aspects of the images and the content that went popular and that did not [12]. The image resolution, color strength, color combinations are some of these. A study on the effect of thin and heavy weight images in Social Media Users [13] gives an idea of the relevance of understanding the usage patterns of different images in SM users. Virality in Social Media is explained using a SPIN framework [14], where, he demonstrates factors like spreadability, propagativity, integration and nexus. A study done by Garimella et. al, [15] show how they used social media image analysis to identify the health related issue of the people in a locality. Certain studies have been done to predict the number of retweets and to obtain values for images in Flickr [16]. Study on the virality of images in Google plus [17] explores the dynamics of the image content that results in virality. Khosla et. al, explains how the images become popular and explains it using two factors namely, image content and social context. Deza & Parikh [19], explains how images go viral from the computer vision perspective. However, besides these studies, there hasn't been much of an exploration in this domain of media information propagation in SM.

3 Methodology

The various steps involved in this study are described below.

1. Collection of twitter data with images – This process includes the identification of twitter profiles with more than 500 followers. The twitter data with images is

collected for a time period of three months ranging from Jan 2018 to March 2018 from these identified celebrity profiles.

2. Cleaning of twitter data. Identifying the tweets with images. Extracting the images to a local path for processing.
3. Performing a descriptive analysis on the extracted tweets to identify the value for various twitter parameters
4. Performing an analysis on images to identify the value for image related parameters like color detection, strength of the color and normalized value based on its strength.
5. Loading all identified features including transactional variables, network features, image features and user related parameters for learning and analysis. This involves slicing and dicing of data to convert it in the format desired for the modelling.
6. Classify the data as training set and testing set. 75% of the tweets and its parameters will be treated as a training set and the rest for testing.
7. Performing various learning techniques namely - Decision Trees, Random Forest, and Neural Networks to analyze the data and identify the relevance of various parameters in making tweets with images go viral. More techniques are used so as to compare the result of the studies.

4 Data Collection

Twitter data was collected from profile ids which had more than 500 followers. 100 profiles were identified as celebrity profiles. Some of the profiles are listed below:

AskNimesh, AustralianSuper, AwesomityFun, AzharuddinKadri, Bayer, Bayer4CropsUS, BayerSuomi, BeShakespeare, BestLoveNotes, BombayTEXT,
Boredwiki, borusanholding, BoschEspana, Bupa, CaIuml5SOS, CapriceHoldings,
Cargill, castawaychild, cheth, CloroxCo, ClubSarcasm, coopuk, Djvasava

Twitter Data was collected from these high profiles for a period of three months starting from January 2018 to March 2018. Python 2.7 was used to extract the data using Twitter APIs. MySQL database was used to store and query for the results. The various parameters extracted include:

tweet_id, source, tweet_text, tweet_favorite_count, tweet_retweet_count, tweet_created_at, listed_count, statuses_count, friends_count, location, favourites_count, name, screen_name, created_at, profile_background_image_url_https, verified, profile_text_color, profile_image_url_https, media, media_url_https, coordinates, hashtags, urls, and retweeted.

4.1 Twitter Data Extraction

Basic building block of modelling exercise is the extraction of tweet (or post) features. The more exhaustive the features are, more accurately we can predict. Feature engineering is considered to have major share in modelling exercise and is therefore the most crucial step. Following section describe various twitter features.

- Tweet_id: Twitter identifies every tweet differently by providing a unique ID; this ID is called tweet_id
- Source: This is the source through which, a user performs an activity in twitter. This could be a mobile app or a web application. It is also identified at the granular level, for e.g. twitter for mac
- Text: Tweet shared or posted by a user on twitter platform comes as text field in the API. All content of a tweet including html links, hashtag, etc. is received as part of this tweet text
- Favorite: Favorite field identifies number of likes received on post. This is an independent variable that we have considered for modelling.
- Retweet: Retweet field identifies number of retweet received on post. This is the second independent variable we have considered for modelling.
- Created at: This is timestamp at which the tweets are created. Timestamp is very important variable for posting on twitter. There are millions of tweets generated every hour, and if tweets are not done in the right timings, that is when your target customers are online, you may miss engagement opportunity.
- Listed: Listed field shows number of times particular person is listed in some other individual’s post. This shows how the demand and popularity of a person/celebrity.
- Statuses: This field shows number of tweets posted by the person till now. This is an aggregate variable which shows activeness of a person on social media platform
- Friends_count: Friends count is the count of people the person is following.
- Location: Location variable shows the places from where the tweets are originating. A diverse location is better for data modelling so as to avoid any kind of bias in the data.
- Favourites_count: This variable shows number of times a person liked tweets of others. This shows social interaction of a person.
- Name: This field shows actual name of a person. A business can use this to identify their target customer like if person is male or female, their religion, region etc. derived from the name.
- Screen_name: Screen name is similar to a user name by which a person is known in Twitter. This not required to be same as the given name.
- created_at: This field shows timestamp on which user created his/her profile on twitter. You can use this variable to calculate age of a person on social media platform
- Media: This field shows, if tweets have any media content attached to it. This is an important field since it helps us to filter tweets with images
- Media_url_https: URL linked with media content
- Retweeted: This indicates if a person shares a post created by his/her own. This it is a very important variable for model building.

From the identified profiles the authors extracted 35, 721 tweets. Mean average of favorites on the shortlisted posts are 101 whereas the mean of text size is 97 characters. Table 1 gives the status of the data collected.

Table 1. Data Status

	avg favorites	friends count	favorites count	text length	listed count	status count
mean	101.92	43912	4406.59	97.09	610.29	14134.91
std	323.92	93852.38	12765.5	37.65	1465.09	23852.17
min	0	0	0	22	0	73
max	1636	705145	187512	318	8980	177862

Location analysis indicates that 17% of the twitter data is from India and rest from other parts of the world. 20% of the profiles did not specify their locations. Tweets were collected from the identified celebrity profiles.

5 Data Analysis

The extraction of tweets and its various other parameters was then followed by the analysis process. Tweets were cleaned to remove noise and other unwanted information. From the downloaded 35,721 tweets only 7,767 tweets were identified to be fit for this study. Other tweets did not have image content attached. Using the Media_url field, images attached to every tweet was downloaded to a local path for faster processing. The tweets with images were then processed for the training purpose. 75% of the data was used for training the model. To identify the success rate of a tweet, following two parameters were used.

- Average likes: This shows the average of favorites received by an individual for his past 200 posts
- Standard deviation of likes: This field shows standard deviation of favorites received by the person in his or her last 200 posts.

If the number of favorites for a tweet exceeds average likes + 2 * standard deviation of likes, then that tweet is considered to be as popular. 75% of the data was labelled as popular and unpopular based on this observation. The image features were then calculated to understand if any of image related parameters are influencing the popularity of tweets.

- Image Feature: For this analysis, we have used image features as a quantitative factor. There are two approaches to it, convert colors into a nominal value; other way is to treat color as a continuous variable. This variable will then give a range of continuous vales that can be better utilized in the model.

All set of variables are then grouped into four different sets

i. Historical variables: favorites_count, listed_count, statuses, average historical likes, and standard deviations of likes; are categorized as historical variables
ii. Transactional variables: Transaction variables are related with the current tweet. Variables like Tweet Creation time (hour), Age of profile (year), tweet_len, retweeted are type of transactional variables.
iii. Network variables: Variables like followers_count and friends_count are counted as network of a person.
iv. Image Features: colors in an image and strength of colors

Different learning approaches where then used to learn and identify the importance of these variables with respect to the popularity of a tweet. Decision Trees, Random Forests and Neural Networks were used to understand this problem.

Decision Trees

Decision Tree is a decision based support tool which is often used as a machine learning approach. A graph that looks like a tree is constructed where each intersection or the node acts as a case for a test. The branches denote the outcome of these tests. It includes chance nodes, decision nodes and end nodes. In this study we uses decision tree to identify if a parameter is really contributing in twitter popularity. This is a supervised learning algorithm. This study have used DecisionTreeClassifier package in python.

Random Forests

Random forest is a supervised learning technique, used for regression, classification, etc. Decisions trees habit of over fitting to the training sets are corrected by this mechanism. It works by fitting number of decision trees by dividing samples, on the overall dataset. Generally, sampling works through bagging technique. Average of each decision tree is taken as the final outcome. RandomForestClassifier in python was used for performing this task

Neural Networks

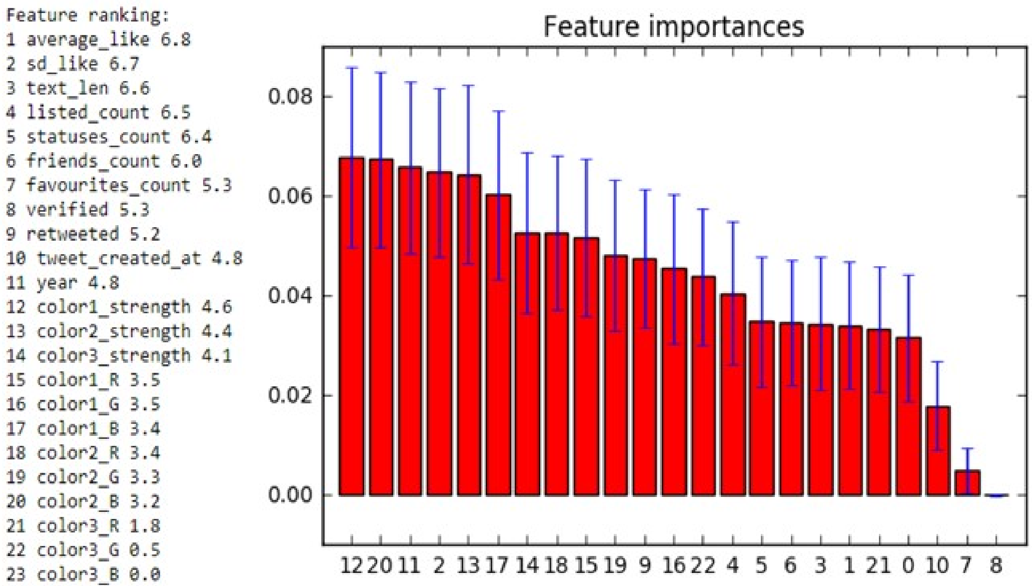
Neural network models are human brain inspired algorithms that are developed to work in the same way our brain process information. Most common used in application form of neural network is multilayer feed-forward network which have several layers, each layer calculate information from its previous layer. Inputs from one node to another are combined through a weighted linear equation. Output is modified based on algorithms before sent as output to the other layer. Weights are generally taken as random values which are passed through several layers. It is recommended to provide pre-learned values to reduce calculation complexity while solving complex problems. MLPClassifier from neural networks in python was used.

6 Results and Discussions

Figure 1 and Figure 2 gives the importance of various features in case of Decision Trees and Random forest separately. It is observed that historical variables contribute more towards popularity. Table 2 compares the result of decision trees and random forests. Table 3, 4 and 5 gives the confusion matrix.

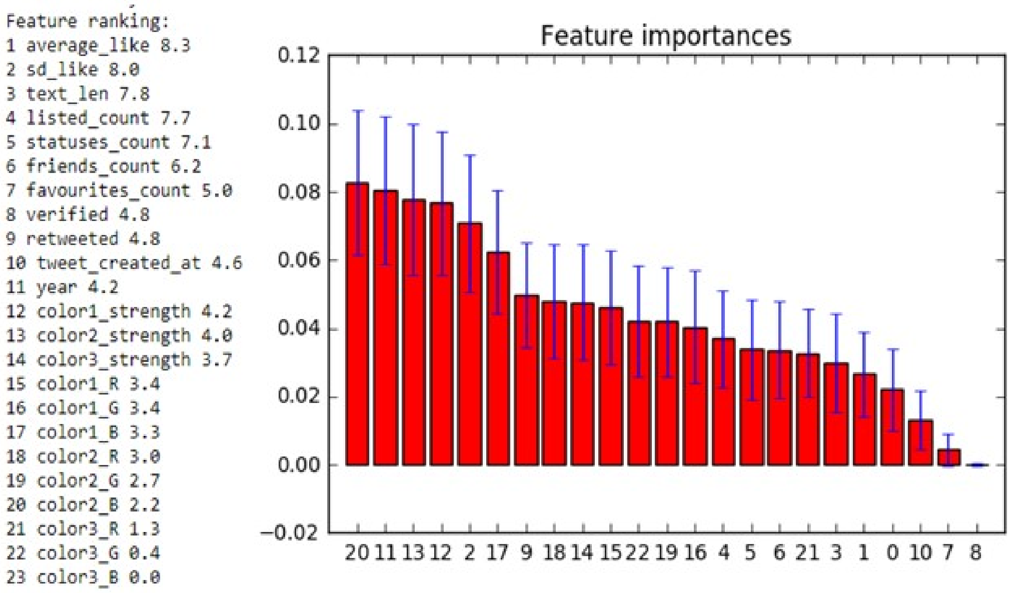
Results of the variable importance shows that how random forests are giving higher weightages to the aggregated and real time variables and lower weightages to the network and image features. As historical variables are the function of network variables, random forest is trying to avoid over fit on this data.

In both algorithms, comparatively a lower importance is given to the network variables and image features. In the second part of analysis on confusion matrix, we have seen higher recall for decision tree with less precision, whereas random forest and neural network have higher precision and lower recall (see Table 6)



```
Feature ranking:
1 average_like 6.8
2 sd_like 6.7
3 text_len 6.6
4 listed_count 6.5
5 statuses_count 6.4
6 friends_count 6.0
7 favourites_count 5.3
8 verified 5.3
9 retweeted 5.2
10 tweet_created_at 4.8
11 year 4.8
12 color1_strength 4.6
13 color2_strength 4.4
14 color3_strength 4.1
15 color1_R 3.5
16 color1_G 3.5
17 color1_B 3.4
18 color2_R 3.4
19 color2_G 3.3
20 color2_B 3.2
21 color3_R 1.8
22 color3_G 0.5
23 color3_B 0.0
```

Fig. 1. Feature importance in case of Decision Trees



```
Feature ranking:
1 average_like 8.3
2 sd_like 8.0
3 text_len 7.8
4 listed_count 7.7
5 statuses_count 7.1
6 friends_count 6.2
7 favourites_count 5.0
8 verified 4.8
9 retweeted 4.8
10 tweet_created_at 4.6
11 year 4.2
12 color1_strength 4.2
13 color2_strength 4.0
14 color3_strength 3.7
15 color1_R 3.4
16 color1_G 3.4
17 color1_B 3.3
18 color2_R 3.0
19 color2_G 2.7
20 color2_B 2.2
21 color3_R 1.3
22 color3_G 0.4
23 color3_B 0.0
```

Fig. 2. Feature importance in case of Random Forest

Table 2: Comparison of Decision Trees and Random Forests

Decision Trees	Random Forests
Historical Variables – 34%	Historical Variables – 37%
Transaction Variable – 26%	Transaction Variable – 26%
Network Variable – 14%	Network Variable – 12%
Image Features – 26%	Image Features -25%

Table 3: Confusion Matrix and Prediction for Decision Trees

Confusion Matrix		Predicted	
Decision Trees		0	1
Actual	0	1820	33
	1	20	37

Table 4: Confusion Matrix and Prediction for Random Trees

Confusion Matrix		Predicted	
Random Forest		0	1
Actual	0	1,832	21
	1	26	31

Table 5: Confusion Matrix and Prediction for Neural Networks

Confusion Matrix		Predicted	
Neural Networks		0	1
Actual	0	1834	19
	1	25	32

Table 6. Precision and Recall

Decision Tree	Random Forest	Neural Network
Precision: 52.9%	Precision: 59.4%	Precision: 62.7%
Recall : 64.9%	Recall : 54.4%	Recall : 56.1%

This study was successful in obtaining, low false negative rates followed by low false positives. Results of confusion matrix is interesting for interpretation and comparisons. Key part of this study is the comparison of results obtained from the decision tree with respect to the random forest method. Decision tree classification is a faster method compared to random forest & neural networks, because it provides a simple interpretation of the dependent variables, which is of interest for many studies.

7 Conclusion

In the light of analysis performed for prediction of tweet popularity based on different factors like - transactional variables, image features, historical variables of the user, and network variables - most important feature obtained was the historical and transactional variables. This behavior has been proven both by decision tree and random forest prediction mechanisms. Image features and network variables did not contribute much to the popularity of the tweet. Increasing the parameters and observations for image and network related parameters might be a problem of future interest. This study will help social media tools to provide capability to automate their process, by helping the businesses and customers to post their content while probability of tweets going viral is comparatively higher. This study was successful in analyzing and predicting the popularity of tweets with images.

References

1. Bakhshi, S., Shamma, D.A. and Gilbert, E. 2014, April. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 965-974). ACM.
2. Pittman, M. and Reich, B. 2016. Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior*, 62, pp.155-167
3. Hu, Y., Manikonda, L. and Kambhampati, S. 2014, June. What We Instagram: A First Analysis of Instagram Photo Content and User Types. In *Icwsm*
4. Kietzmann, J.H., Hermkens, K., McCarthy, I.P. and Silvestre, B.S., 2011. Social media? Get serious! Understanding the functional building blocks of social media. *Business horizons*, 54(3), pp.241-251
5. Gupta, A., Lamba, H., Kumaraguru, P. and Joshi, A. 2013, May. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In Proceedings of the 22nd international conference on *World Wide Web* (pp. 729-736). ACM.
6. Toubia, O. and Stephen, A.T. 2013. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter?. *Marketing Science*, 32(3), pp.368-392
7. Joseph, N., Kar, A.K., Ilavarasan, P.V. and Ganesh, S. 2017. Review of discussions on internet of things (IoT): Insights from twitter analytics. *Journal of Global Information Management* (JGIM), 25(2), pp.38-51
8. Russell, M.A. 2013. Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More. " *O'Reilly Media, Inc*
9. Abel, F., Gao, Q., Houben, G.J. and Tao, K. 2011, July. Analyzing user modeling on twitter for personalized news recommendations. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 1-12). Springer, Berlin, Heidelberg
10. Kotsiantis, S.B., Zaharakis, I. and Pintelas, P. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, pp.3-24
11. Mangold, W.G. and Faulds, D.J. 2009. Social media: The new hybrid element of the promotion mix. *Business horizons*, 52(4), pp.357-365.
12. Vinner, S. and Dreyfus, T., 1989. Images and definitions for the concept of function. *Journal for research in mathematics education*, pp.356-366.

13. Smeesters, D., Mussweiler, T. and Mandel, N. 2009. The effects of thin and heavy media images on overweight and underweight consumers: Social comparison processes and behavioral implications. *Journal of Consumer Research*, *36*(6), pp.930-949
14. Mills, A.J. 2012. Virality in social media: the SPIN framework. *Journal of public affairs*, *12*(2), pp.162-169
15. Garimella, V.R.K., Alfayad, A. and Weber, I. 2016, May. Social media image analysis for public health. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5543-5547). ACM
16. Van House, N.A. 2007, April. Flickr and public image-sharing: distant closeness and photo exhibition. In *CHI'07 extended abstracts on Human factors in computing systems* (pp. 2717-2722). ACM
17. Guerini, M., Strapparava, C. and Özbal, G. 2011, July. *Exploring Text Virality in Social Networks*. In *ICWSM*
18. Wu, B., Cheng, W.H., Zhang, Y. and Mei, T. 2016, October. Time matters: Multi-scale temporalization of social media popularity. In *Proceedings of the 2016 ACM on Multimedia Conference* (pp. 1336-1344). ACM
19. Deza, A. and Parikh, D. 2015. Understanding image virality. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1818-1826).