Predicting Likes and Retweets on Twitter

Jared Wright

Spring 2019

Brigham Young University

Predicting Likes and Retweets on Twitter

Abstract

In this paper I examine what words and factors in a tweet promote "likes" and retweets on Twitter. I use data pulled from Twitter's API on President Trump. I determine what factors lead to more likes and retweets. For example, I determine the impact of including a link on likes and shares.

Introduction

Social media has become an integral part of American life. It influences our perceptions of the world around us. For many Americans social media is a primary news source. However, whether or not someone sees a post in their feed is in part dependent on the amount of likes and shares the post received. Thus a better understanding of what entices people to like and share posts would enable individuals and businesses to reach a wider audience. In this data project I attempt to identify what words persuade people to like and share.

A prime opportunity for study comes from President Trump. He frequently uses Twitter as a primary means of communication. Additionally, he uses simple language. He tends to repeat key words and phrases over time. This repetition allows us to identify whether these key words motivate the audience to like and share his tweets. Finally, he has a large base of followers, and each of his many tweets receive thousands of favorites and retweets. All of this makes for a large and clean sample. For this reason, data on President Trump's tweets is well suited for this study.

A lot of literature already discusses how President Trump's tweets impact the political scene. For example, Yu Wang et. al. as well as Ramona Kreis examine Trump's communication style on the political atmosphere. Wang et. al. find that Trump followers "like" tweets more when he references Democrats. Kreis examines how Trump's tweets affect right-wing populist discourse.

Other literature deals with analyzing what promotes sharing on social media. Stefan Stieglitz and Linh Dang-Xuan conclude that emotionally charged tweets are more likely to be retweeted. Danah Boyd et. al. study how retweeting mimics a conversation. Additionally, Macskassy and Michelson attempt to identify what types of information are being retweeted.

My hypothesis is that my findings will support the findings of Stieglitz and Dang-Xuan as well as Wang et. al. Emotionally charged words and indicators seem more likely to generate an emotional response in the reader, which will induce them to like and retweet the post. Additionally, I think that attacking opponents will endear President Trump to followers. Thus referencing Democrats should garner more likes and retweets.

Dataset Description

Notably, the findings for President Trump will not be completely generalizable to all social media users. However, findings will probably be generalizable to Twitter users and to users with lots of followers. Because of the broad nature of this question, this data project cannot completely answer the question of interest. But the data analysis will not be entirely in vain. This analysis will be one part of the answer needed to complete the puzzle.

I downloaded the data from Twitter's API. The dataset contains 3,200 tweets from President Trump's twitter account, @realDonaldTrump. Each tweet is an observation in the dataset. Along with the full text of each tweet, each observation contains a date/time stamp, information about the tweet, and information about the user.

Data Modification

Because the dataset contained only one user, I dropped all the user variables. However, I carefully went through the variables that related information about the tweet, and dropped what would not be valuable in the analysis. I dropped tweet information variables such as latitude and longitude at which the tweet was made (this information was disabled for President Trump), the language of the tweet, and the users and hashtags mentioned in the tweet. Numerous other variables I kept, such as an indicator if the tweet was a retweet, a unique tweet id, the text of the tweet, and the number of likes and retweets a given tweet received.

Not all of the tweets in the dataset were valid for analysis. This is because some of the tweets were retweets. Retweets are not included in the analysis because they do not contain the unique wording of President Trump in general. Additionally, retweets theoretically would have more likes and shares because they reach a wider audience base. Retweets reach the audience base of both President Trump and the original Tweeter; whereas original tweets from President Trump only begin in his audience base.

About 800 of the tweets are retweets. Hence the final dataset contains 2,610 observations. Despite the fact that the dataset is not as large as it could be, it is sufficiently large to run conclusive tests as well as preliminary analysis.

However, the primary limitation of the dataset is not the size, but the relatively focused scope. The data will be perfect for revealing what key words motivate retweets and likes for President Trump on Twitter. Our question is more general, though. Only limited aspects of the analysis may be generalizable.

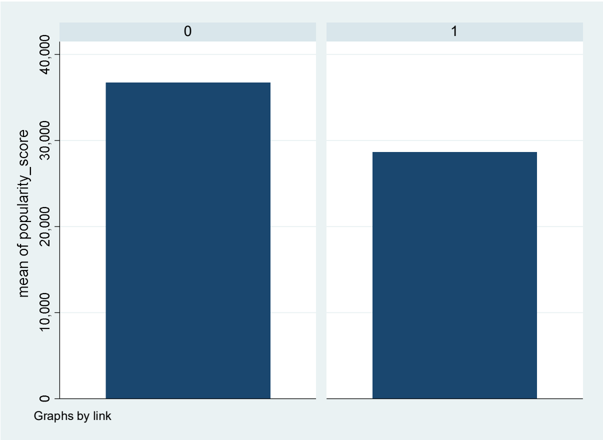
Descriptive Statistics

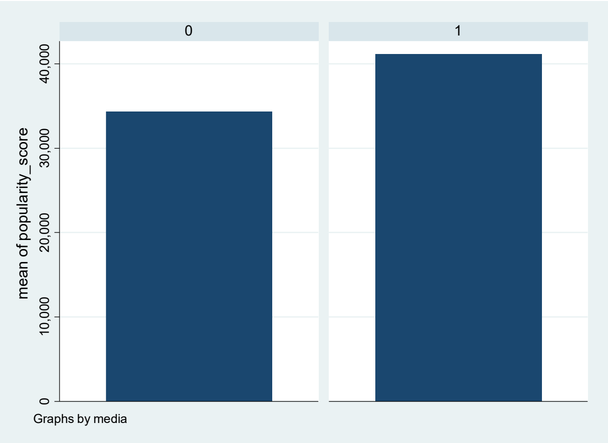
It is hard to visualize summary statistics outside a regression because of the nature of the binary variables. Nevertheless, below are a few bar charts showing the effect of a few variables on the popularity of a tweet. I created a composite score for popularity that weights likes and shares equally. One retweet is given less weight than one share, but shares account for 50% of the weighting and retweets account for the other 50%.

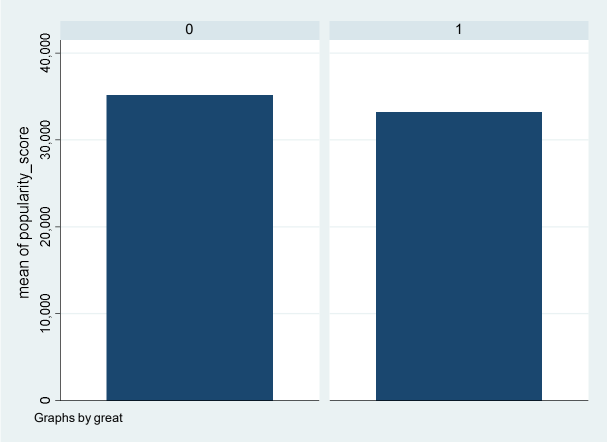
In the first chart, link has a score of one if President Trump included a link in his tweet. As is apparent, tweets with links are shared and liked less than tweets that do not include a link. The next graph is sorted by if President Trump included the word “media” in his tweet. The next one after measures the effect of President Trump using the word “great”. Finally, the last chart measures the effect of using an exclamation point.

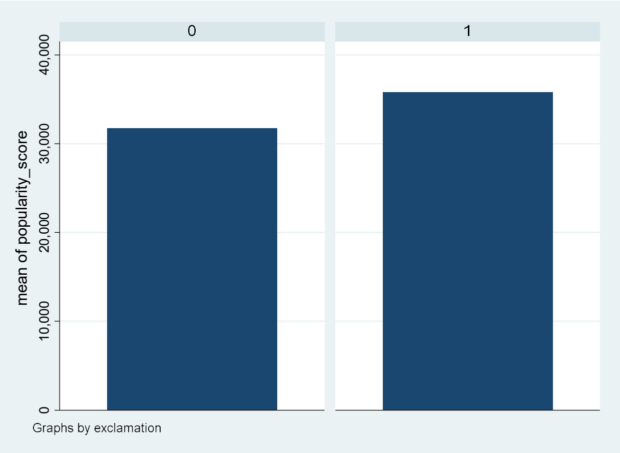
These are just four examples of the many variables I included in the test. Some variables, such as including a link, the word “great”, tweeting at another individual, or the words “China” and “thank”, have a negative correlation with this popularity score. Other words, such as “media”, “wall”, “border”, “America”, and “I”, have a positive correlation with the popularity of a tweet.

Ultimately, this preliminary data does not tell us much about whether or not these relationships are statistically significant. However, it does give us an idea of what things may be important in gaining retweets and likes of a post.









Analysis

Table 1 presents the results of the final regression analysis. I ran a regression with more variables, but ended up dropping the variables that were not anywhere close to statistically significant. Notice that many variables were statistically significant, especially in the regression on weighted popularity score.

Words in quotation marks in the "Word or device" column of the regression table are words included in the text of President Trump's tweets. Italicized words come not from the text, but from other devices used in the tweet, which I outline here. *Exclamation* is an indicator variable indicating whether or not President Trump included an exclamation point in his tweet. *Link* indicates a link included in the tweet. *Mention* indicates whether Trump mentioned someone in the tweet. (A user mentions another user by including the "@username" handle). *Self-reference* indicates whether or not Trump referenced himself using words "I" or "me". Finally, *Constant* is a constant variable that indicates the base number of likes and retweets.

Table 1

Impact of Linguistic Devices on Likes and Retweets

Dependent Variable	Number of Likes	Number of Retweets	Weighted Popularity Score
Word or device	Estimates (SE)	Estimates (SE)	Estimates (SE)
“America”	23.992**	57.64***	2355.455**
	(6.339)	(6.341)	(960.9938)
exclamation	-25.85***	-25.49***	3399.465***
	(4.942)	(4.945)	(664.3813)
link	-19.07***	-19.56***	-7068.819***
	(3.833)	(3.845)	(710.5097)
“great”	-9.320**	-13.61***	-2222.102***
	(4.128)	(4.024)	(687.1439)
Mention	21.08***	21.09***	-7888.987***
	(6.145)	(6.148)	(739.2107)
“China”	9.097*	11.38**	-7690.073***
	(4.910)	(4.895)	(1587.667)
“wall”	78.58***	79.81***	5777.495***
	(4.699)	(4.696)	(1517.695)
Self-reference	-42.00***	-40.67***	1605.118**
	(4.310)	(4.301)	(671.2807)
“President”	-19.41***	-21.19***	3346.839***
	(4.036)	(4.025)	(917.0234)
“media”	3.669	-40.67***	4370.831***
	(4.606)	(4.301)	(1374.396)
“people”	15.76***	-21.19***	2011.735**
	(4.844)	(4.025)	(845.6365)
“MAGA”	-13.44**	-13.69**	-5514.741***
	(6.254)	(6.257)	(1601.87)
“Russia”	5.119	5.016	1613.891
	(5.149)	(5.152)	(1463.386)
“Democrats”	0.672	1.104	-1384.109
	(5.604)	(5.607)	(1482.993)
“Korea”	8.141	7.656	-3063.025
	(8.053)	(8.058)	(2313.105)
“Trade”	33.30***	32.93***	1286.199
	(5.906)	(5.909)	(1789.846)
“New”	68.90***	69.58***	-1989.268**
	(5.633)	(5.636)	(932.6198)
“Fake news”	-47.42***	-47.49***	3554.773**
	(5.222)	(5.224)	(1642.845)
Constant	-4.555	-4.409	35902.3***
	(5.223)	(5.225)	(711.3943)
Observations	2,563	2,563	2,563
Adjusted R-squared	0.1628	0.1216	0.1429

Standard Errors in Parentheses *Significant at 0.10 **Significant at 0.05 ***Significant at 0.0

References

Wang, Y., Luo, J., Niemi, R., Li, Y., & Hu, T. (2016). In International AAAI Conference on Web and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ ICWSM/ICWSM16/paper/view/13054/12839

Kreis, Ramona. (2017). “The ‘Tweet Politics’ of President Trump”. Retrieved from https://doi.org/10.1075/jlp.17032.kre

Lee, K.; Mahmud, J.; Chen, J.; and Zhou, M. 2015. Who will retweet this? detecting strangers from twitter to retweet information. ACM Transactions on Intelligent Systems and Technology 4:1–25.

Sofus A. Macskassy and Matthew Michelson. 2011. Why do People Retweet? Anti-Homophily Wins the Day!. In ICWSM.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone’s an Influencer: Quantifying Influence on Twitter. In WSDM.

Stieglitz, Stefan; and Dang-Xuan, Linh. (2014). Emotions and Information Diffusion in Social Media—Sentiment of Microblogs and Sharing Behavior. Journal of Management Information Systems. https://doi.org/10.2753/MIS0742-1222290408

Boyd, D.; Golder, S.; and Lotan, G. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. In R.H. Sprague (ed.), Proceedings of the 43rd Annual Hawaii International Conference on System Sciences. Los Alamitos, CA: IEEE Computer Society, 2010.