Causal Loop

MONDAY, FEBRUARY 13, 2012

Tweets vs. Likes? An Analysis of the Monkey Cage

A while back, Joshua Tucker issued a challenge on The Monkey Cage:

> Here at The Monkey Cage we allow people to "Tweet" posts to their Twitter followers, and "Like" posts to their Facebook friends. Lately I've noticed that some posts get more tweets than likes, some get more likes than tweets, and others get roughly the same amount. Anyone have any idea why?

Challenge accepted.

I was actually surprised to find that this question has already been looked at by other data science bloggers. A quick google search for "Tweets vs. Likes" got me to Edwin Chen's blog where he posed the exact same question as Joshua did:

> It always strikes me as curious that some posts get a lot of love on Twitter, while others get many more shares on Facebook:
>
> What accounts for this difference? Some of it is surely site-dependent: maybe one blogger has a Facebook page but not a Twitter account, while another has these roles reversed. But even on sites maintained by a single author, tweet-to-likes ratios can vary widely from post to post.
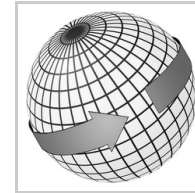
He analyzes the data from a few tech-related blogs, comparing the tweet-to-like ratio for each post to various post attributes and finds that:

> tl;dr Twitter is still for the techies: articles where the number of tweets greatly outnumber FB likes tend to revolve around software companies and programming. Facebook, on the other hand, appeals to everyone else: yeah, to the masses, and to non-software technical folks in general as well.

This nerd/normal divide corresponds surprisingly well to Joshua's initial set of hypotheses.

Humor vs. wonkishness hypothesis: The funnier a post, the more likely it is to go on Facebook; the wonkier the post, the more likely it is

to get tweeted.

The graphics hypothesis: The more graphics, the more likely it is to go to Facebook. The more text, the more likely it is to be tweeted.

The source of visitors hypothesis: Visitors outside academia are more likely to post to Facebook; academics who read blogs are more likely to tweet.

Is this really the case? To obtain the actual data, I wrote a quick screen scraping script and went through all posts from this February up until about May of last year. At some point after that, no likes or tweets appear to be recorded for most of the posts. In total, I scraped around 860 posts, 492 of which had both tweets and likes.

I use a modified version of Edwin Chen's tweet-to-like ratio as the dependent variable. In order to avoid dividing by zero since many posts have only tweets and no likes, I add 1 to the quantity of both tweets and likes for a given post. I then take the base-10 log of the modified tweet/like ratio to linearize the dependent variable for regression analysis. For brevity, let's call this measure the "tweet rating" - positive values indicate more tweets than likes while negative values indicate more likes than tweets.

Since the third of Joshua's hypotheses is untestable with the data that I could obtain, I'll focus on the first two. Graphics and length are directly measurable. I use a dummy variable indicating whether or not a post includes a graphic (i.e. img tags) and another indicating whether a post has an embedded video. For length, I use only a basic word count measure. Since this may not capture the "complexity" of a post well, I also include the Flesch-Kinkaid grade level (a rather rough measure, but the best quantitative one that I could come up with quickly).

Wonkiness vs. Humor is a bit harder to capture. While it would be interesting to do a full analysis of each post to determine the sentiment (using something like Sentiwordnet and a natural language processor), I simply don't have the time. As a proxy, I use post categories. A lot of the categories are rather neutral but a few stand out as relevant in the nerd/normal framework. Frivolity and especially the Ted McCagg Cartoons are definitely more humor-oriented. Conversely, I found the "Data," "Academia," "Methodology," and "IT and Politics" categories more "wonky" than the rest. Each is coded as a 0-1 dummy variable.

Causal Loop: Tweets vs. Likes? An Analysis of the Monkey Cage

2/17/12 9:43 AM

What's the relationship? Tables 1 and 2 (at the bottom of the post) give the results of 8 regressions with different combinations of the above variables. The results appear to be strongly supportive of Joshua's hypotheses. First, the presence of some sort of image has a highly significant negative effect ($p < .01$) on the tweet rating. Put visually:
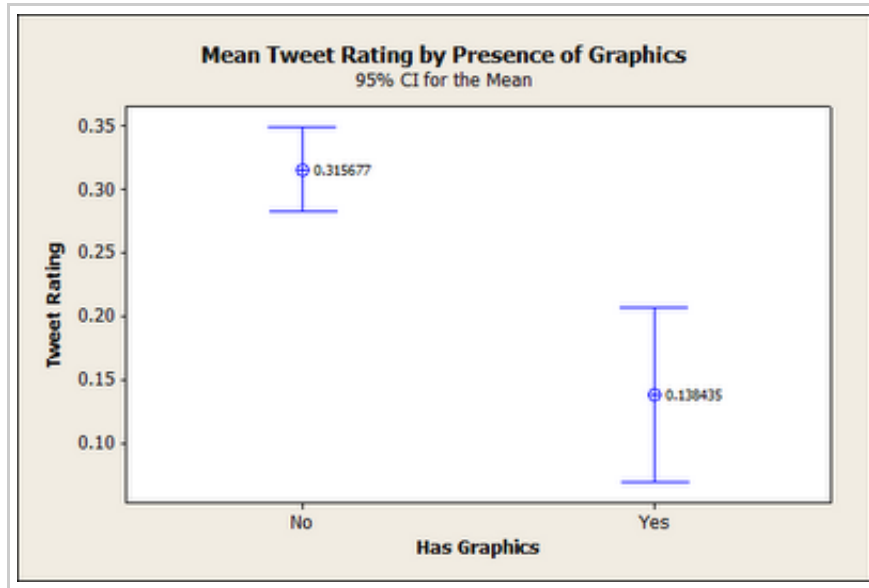
Posts with graphics therefore tend to have a lower tweet/like ratio, indicating more popularity on Facebook. Length and Grade Level, however, do not appear to be significantly associated with the Tweet Rating (perhaps a more nuanced measure of wonkiness would be appropriate).

The relationship between the categories and tweet factor also appear to support the nerd/normal story. When considered alone, the Frivolity and TMC Cartoon categories also have a significant negative effect on the tweet rating. Conversely, the Data and IT and Politics categories are associated with a higher tweet rating (although the Data category is only significant at the < .10 level).

This is certainly a very rough look at the data, but it does seem to suggest that Joshua's hypotheses have some validity. Posts that have graphics or are funny are more "likable" while wonkier posts are more "tweetable." I would add that the first relationship is a bit stronger than the second since its difficult to find a good measure of "wonkiness" (especially since almost *all* posts on the Monkey Cage are relatively wonky). That the more "tech" categories (Data and IT/Politics) had a positive effect on the Tweet Rating might lend support to Edwin Chen's argument that the Twitter ecosystem is geared specifically towards technology nerds.

Causal Loop: Tweets vs. Likes? An Analysis of the Monkey Cage

2/17/12 9:43 AM

I'm not sure that the results say much about the composition of Facebook vs. Twitter - a lot of people use both. I do, however, think that they may hint at a key difference in the content-sharing incentives behind both services. Facebook is much more graphically-oriented than Twitter. The new "Timeline" profile structure makes this absolutely clear. Moreover, pictures and videos receive much more visual prominence in a user's Facebook feed than simple text. Therefore, there is a much greater chance that shared content will be noticed if it contains an eye-catching photo or graphic. Conversely, Twitter feeds are pure text, which means that graphics are not a means of distinguishing one's tweets from those of others. The value of graphics is greater on Facebook, which gives users a strong incentive to share content that has some visual component in order to get noticed.

Certainly there are other possible explanations. Tweets tend to be more public than Facebook posts, which are aimed more at one's circle of known friends and acquaintances. Even if the bases of users for both services is similar, there may be a difference in the types of people who prefer to use Twitter vs. those who prefer Facebook (nerds vs. "normal people"?)

It would be interesting to extend this to other political blogs to see if the relationship holds. Would policy-oriented blogs be different from academic blogs? Are certain authors more tweet-prone or like-prone? Interesting questions for anyone who has an excess of free time.

You can download the data I used for this post in STATA format or in a tab-delimited text file.

I obtained the data using a screen scraper written in Python. You can download it here along with the conversion script to make the results usable in statistical software. Hat tip to Ed Cranford for his Flesch-Kinkaid score calculator. Required additional libraries are nltk, PAMIE and BeautifulSoup. I also have the full texts for each post in the data set which are available via e-mail request.

Appendix - Tables

Table 1 - Standard OLS regression of "Tweet Factor" on independent variables

Independent Variable	1	2	3	4
	-		-	

Graphics	0.1772***	-0.1760***	0.1702***	
	(-5.05)	(-5.01)	(-4.80)	
Length		-0.000032		
		(-1.12)		
Grade Level		0.003513		
		(0.90)		
Frivolity			-0.1063	-0.1687**
			(-1.27)	(-2.02)
TMC Cartoon				
Data				
IT and Politics				
Methodology				
Academia				
Constant	0.3157***	0.2925***	0.2768***	0.2758***
	(17.68)	(6.88)	(17.47)	(17.40)
Num. Observations	860	860	860	860

*T-values in parentheses. Significant at: * = .10, ** = .05, *** < .01*

Table 2 - Standard OLS regression of "Tweet Factor" on independent variables cont.

Independent Variable	5	6	7	8
Graphics		-0.1798***	-0.1797***	-0.1843***
		(-5.13)	(-5.11)	(-5.25)
Length				
Grade Level				
Frivolity				
TMC Cartoon	-0.3588**			
	(-2.21)			
Data		0.2290*	0.2293*	0.2356*
		(1.75)	(1.75)	(1.80)
IT and Politics				0.2723**
				(2.40)
Methodology			0.0476	0.0528
			(0.55)	(0.61)
Academia			-0.0254	-0.0201

			(-0.36)	(-0.29)
Constant	0.2731***	0.3132***	0.3128***	0.3084***
	(17.48)	(17.50)	(16.79)	(16.52)
Num. Observations	860	860	860	860

*T-values in parentheses. Significant at: * = .10, ** = .05, *** < .01*

POSTED BY ANTON STREZHNEV AT 1:17 PM

1 person +1'd this

LABELS: FRIVOLOUS, STATISTICS, THE MONKEY CAGE

---

0 COMMENTS:

POST A COMMENT

Enter your comment...

Comment as: Select profile...

Publish   Preview

Home   Older Post

Causal Loop: Tweets vs. Likes? An Analysis of the Monkey Cage

2/17/12 9:43 AM

Subscribe to: Post Comments (Atom)