



Contents lists available at ScienceDirect

IJRM

International Journal of Research in Marketing

journal homepage: www.elsevier.com/locate/ijresmar

Full Length Article

Composing tweets to increase retweets

Nima Y. Jalali^a, Purushottam Papatla^{b,*}^a Department of Marketing, Belk College of Business, University of North Carolina-Charlotte, 9201 University City Blvd., Charlotte, NC 28223, United States of America^b Sheldon B. Lubar School of Business Administration, University of Wisconsin-Milwaukee, Milwaukee, WI 53211, United States of America

ARTICLE INFO

Article history:

First received on December 15, 2017 and was under review for 9 months
Available online 12 July 2019

Area Editor: Andrew Stephen

Keywords:

Retweets
Sales promotions
Reading theory
Psycholinguistics
Poisson mixtures
Hierarchical Bayes

ABSTRACT

Brands tweet not only to communicate with followers but also to reach large audiences rapidly when the tweets are retweeted by the followers. People however will retweet only if they recognize within a few seconds that the tweet is on an interesting topic. Brands therefore need insights into how to compose tweets to facilitate topic recognition even when they are just scanned. This is the issue that we address in this research. Specifically, drawing on findings in psycholinguistics, we empirically investigate if tweets composed such that they include more topic-related words that are located closer to the start get more retweets. Results from an investigation of sales-promotional tweets by sixty brands in four categories indicate that tweets that are composed as above do get more retweets. We repeat the investigation using tweets on several other topics from a natural experiment that generated pairs of tweets where each pair is on the same topic but each tweet in the pair is composed differently. This investigation reconfirms the findings from the analysis of retweets of sales-promotional tweets. We conclude by presenting an approach for how social media managers can compose tweets based on our findings.

© 2019 Elsevier B.V. All rights reserved.

1. Introduction

Twitter has become an important channel for brands to communicate with customers and several brands including most of the Fortune 500 are very active on the network (Barnes, Lescault, & Homes, 2015). Starbucks, for instance, has sent more than one hundred and thirty six thousand tweets¹ to about twelve million followers. An advantage of tweets is that, if retweeted, their message can reach large audiences at no cost to firms. For instance, 13 tweets by the automobile brand Mazda in January were retweeted 449 times by the brand's followers who themselves had 447,399 followers.² Similarly, a tweet (Fig. 1) by the restaurant chain Arby's to the singer Pharrell Williams was retweeted about 78 thousand times (Advertising Week, 2014a). The tweet and retweets were viewed about 160 million times on social media and the Internet (Advertising Week, 2014b) thus presenting the brand as many times to audiences. Such increased exposure can lead to higher sales (The Wall Street Journal, 2017) and monetary benefits in several categories. For instance, causal investigations have shown that they can increase product sales (Kumar, Bhaskaran, Mirchandani, & Shah, 2013), stimulate contributions to political campaigns (Petrova et al., 2016), and increase viewer-

* Corresponding author.

E-mail addresses: nima.jalali@uncc.edu, (N.Y. Jalali), papatla@uwm.edu. (P. Papatla).¹ <https://twitter.com/starbucks>² We computed this by collecting data on the number of followers of those who retweeted.



Fig. 1. Arby's Tweet about Singer Pharrell Williams and Grammy's.

ship of television shows (Gong, Zhang, Zhao, & Jiang, 2017). More broadly, the literature (Stephen and Galak, 2012) also provides causal evidence that firm generated content on social media can increase product sales.

Retweets, however, can happen only if the tweets are read and recognized by recipients to be on topics that they would like to tweet about. People, on the other hand, may devote only few seconds to reading a tweet because they spend no more than a few minutes per day on Twitter (Advertising Week, 2014c). Tweets, therefore, have to be composed so as to attract attention and facilitate recognition of their topics even if they are read rapidly and cursorily. Despite the increasing interest in Twitter as a medium (Malhotra, Malhotra, & See, 2012) and in the role of tweets' content on retweets (Araujo, Neijens, & Vliegenthart, 2015), there are few insights in the literature into this.

The few recommendations in the literature focus on attracting attention to a tweet rather than increasing the likelihood that its topic is recognized. Malhotra et al. (2012) for instance recommend that brands should compose tweets to followers so as to attract attention that will "make them read" the tweet. The specific approach they recommend is to include words like "WOW", "LOOK" or "TODAY ONLY!" (p. 64) at the start of the tweet. Their approach thus relies on surprising and creating a sense of urgency to attract attention and reading. Recent findings (Zhang, Moe, & Schweidel, 2016) however demonstrate that retweets are driven more by whether people see a fit between a tweet's topic and what they like to tweet about. The challenge for brands, however, is the little time that people spend on Twitter and individual tweets. As a result, a tweet may not even be read fully and people may not realize that it is on a topic that they like to share by retweeting. There is therefore a real need for brands' social media managers to understand how to compose a tweet to help followers recognize its topic even if it is only scanned and not read fully. This is the gap that we address in this research.

Our investigation is based on findings in the psycholinguistics literature that, when people do not put much effort into reading, they may just comprehend a sentence by scanning a subset of words (Rayner, Alexander, Jane, & Charles Jr, 2012). In the context of tweets therefore people may not retweet unless, early in the scanning process, they see topic-related words (Griffiths, Steyvers, & Tenenbaum, 2007) that lead them to conclude that the tweet is on a topic of interest. For instance, followers interested in social interactions (Lovett, Peres, & Shachar, 2013) may only retweet the Arby's tweet if they see the word Grammys – a topic that can stimulate conversations about the awards with friends. The tweet may therefore have to be composed such that words related to the Grammy's are likely to be spotted soon after scanning begins so that the reader does not lose interest in reading further. Additionally, because comprehension requires a subset of words to be scanned, additional words that are semantically related to the Grammy's (Griffiths et al., 2007) may have to be included at different points in the tweet. This would increase the likelihood that they are also scanned and the topic is comprehended. Our research is aimed at investigating whether composing tweets based on such principles can increase retweets. Specifically, we investigate whether locating topic-related words earlier in the tweet and increasing the number of those words in the tweet can increase retweets.

We conduct our investigation in three phases. In phase 1, we focus on sales promotional tweets (SPTs) among the tweets by sixty-two brands in four categories and look for *prima facie* evidence that early location of and higher space devoted to topic-related words in tweets affect retweets. Data for this includes tweets posted over a fifteen month period between February 4th, 2011 and April 26th, 2012, by 62 brands in the automobile, food and beverage, dining, and airline categories. Based on evidence from this phase, we use propensity score matching methods (Rosenbaum & Rubin, 1983) in a second phase to investigate whether early location of and space devoted to topic-related words in these tweets *cause* retweets. Results from this phase provide evidence of causality. Finally, in phase three, we use a subset of a large corpus of about 1.77 million topic-author controlled tweets that arise from a natural experiment on Twitter (Tan, Lee, & Pang, 2014) to rule out endogeneity and investigate the generalizability of the causal relationship identified in phase 2. Each tweet in this data has a pair on the same topic and by the same author but the two tweets differ in their composition. Further, most authors of the tweets in this dataset are individuals and the tweets cover thousands of topics. For instance, a sample of 18,406 tweets that we investigate from this data included 4671 unique topics following the "#" symbol in the tweets. Collectively, these features allow us to reliably address the issues of endogeneity and generalizability. Specifically, because the tweets are by individuals rather than brands, brand managers' expertise in composing SPTs to increase retweets as an alternative explanation for our findings in phases 1 and 2 can be ruled out if our findings are replicated with this data. Additionally, because tweets by individuals are likely to be on topics other than what brands are likely to tweet about and also because of the large number of unique topics, replication of our findings from the first two phases in this phase will also demonstrate the generalizability of our findings.

We choose sales promotions as the topic for the phase 1 investigation because it offers both theoretical and managerial advantages. Theoretically, the literature on topic models (Griffiths et al., 2007) indicates that topics are categorized in human memory as a set of frequently occurring topic-related words. Given its extensive use by businesses, and the frequent use of words and phrases like “free” and “free gift”, sales promotions are likely to be categorized as a topic consisting of such words. Tweets on sales promotions which include these frequently used words would therefore help us investigate whether early location of and more space devoted to topic-related words increase retweets. Managerially, as well, analyzing retweets of SPT’s is useful because our findings can increase the response to the promotions. Additionally, if the retweets reach people who do not follow the brand but follow its followers, their interest in the brand may be stimulated converting them also into followers of the brand. This can increase future sales (The Wall Street Journal, 2017; Twitter, 2018).³

There have been other investigations of retweets in the literature. For instance, Malhotra et al. (2012) develop recommendations for how to increase retweets based on a descriptive analysis of a sample of 1150 tweets. Zhang et al. (2016) also investigate retweets but focus on how the fit between a tweet’s topic and the audience’s retweeting interests plays a role. Our research thus adds to these investigations. Specifically, the primary contribution of our research is in demonstrating that the locations of and space devoted to topic-related words in tweets affect retweets. We also demonstrate the business value of our findings by collecting and investigating additional data on how retweets affect the number of followers of the brands that we study in phases 1 and 2. Findings from this investigation show that an increase in retweets also increases the number of followers of the brands. Because an increase in the number of followers also increases the number of retweets (Araujo et al., 2015), increasing retweets through better composition of tweets should have positive dynamic effects on future retweets and hence the business benefits of retweets (Gong et al., 2017, Kumar et al., 2013, Petrova et al., 2016) such as sales (The Wall Street Journal, 2017; Twitter, 2018).

In the next section, we briefly discuss previous research on retweets and the related broader field of message transmission in online networks, and the findings in the reading and psycholinguistics literatures on how people read text. We follow this with a description of the data we use in phases 1 and 2 along with operational definitions of variables and follow with a description of our model and empirical results from phase 1. Following this, we present results from our investigations in phases 2 and 3. The paper concludes with a section that empirically demonstrates that retweets of brands’ tweets increase the number of followers, a discussion of the contributions and limitations of our work, and directions for future research.

2. Literature

2.1. Research on retweets

Malhotra et al. (2012) were one of the early investigators of retweets and find that they are higher for shorter tweets and tweets that include the request to retweet #RT_if. Additionally, they find that sales promotional tweets that include time-sensitive announcements (e.g., “... Friday, March 4. Stay tuned for details”) get more retweets.

Subsequent investigations by Araujo et al. (2015) provide additional insights. Their analysis of over four hundred thousand retweets of nineteen thousand tweets by one hundred brands shows that including links to photos and videos and characters like the hashtag increase retweets. The positive effect of the hashtag is important in the context of our research because the word that follows it identifies the tweet’s topic. Thus, Araujo et al.’s (2015) research also indicates that clearly defining and highlighting the tweet’s topic with the hashtag increases retweets. Additionally, they find that tweets posted on Tuesday through Thursday receive more retweets than those posted on other days of the week.

More recently, Zhang et al. (2016) also focus on the role of tweets’ topics in retweets. Their goal is to investigate whether the extent of fit between tweets’ topics and topics that followers themselves like to tweet about affects the number of retweets. Fit is computed using Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003). Zhang et al.’s (2016) results suggest that the closer a tweet’s topic(s) are to the topic(s) that followers themselves tweet about, the more retweets it is likely to get. They therefore recommend that brands should compose tweets using topic-related words for topics of interest to followers. They however do not address the questions of whether the location of and space devoted to topic-related words in these tweets affect retweets.

2.2. Message transmission in online networks

A different but related research stream on why people propagate messages in online networks has also been growing in the literature. This research is also relevant to our work because it can provide insights into factors that increase the diffusion (Stephen & Lehmann, 2016) of messages (or tweets in our case). Some examples of research into the transmission of messages in online networks include Chae, Stephen, Bart, and Yao (2016), Stephen and Lehmann (2016), and Zubcsek and Sarvary (2011).

Chae et al. (2016) investigate message propagation in seeded marketing campaigns or SMC’s. Their specific goals are understanding whether messages sent by influencers selected by firms to write about specific products on social media stimulate

³ We also empirically demonstrate the effects of retweets in attracting new followers to brands in the Managerial Implications section.

discussion of those and related products among the non-seeded population. Their study of 39 SMC's finds that word of mouth about the product does spread but spills over into word of mouth about other products from the firms as well as about the entire product category of the seeded products. In our context, however, Chae et al.'s (2016) findings suggest that tweets by influential brands that have more followers than other brands, or more fans on social media platforms like Facebook, are more likely to be retweeted. Our empirical results, in fact, confirm this.

Stephen and Lehmann (2016) also explore the benefits to firms of careful selection of the recipients of their messages. Specifically, they focus on whether sending messages to recipients who are connected to a large number of people, those with greater connectivity (Hinz, Skiera, Barrot, & Becker, 2011) would speed the diffusion of the message. Their findings indicate that careful selection and targeting of high social connectivity recipients does speed the transmission of the message. While we do not investigate the speed with which firms' tweets are retweeted, Stephen and Lehmann's (2016) findings also provide indirect evidence of how having a large number of followers can provide retweeting benefits to brands.

2.3. How people read

Psycholinguists have studied how people read and comprehend text for over a century (Dearborn, 1906, Frazier & Rayner, 1982, Inhoff & Weger, 2005, Irwin, 1992, Kambe, Rayner, & Duffy, 2001, Matin, 1974, Rayner, 1998, Rayner et al., 2012, Townsend & Bever, 2001). Findings suggest that reading and comprehension of a sentence of text happens sequentially in two stages. The reading stage itself proceeds in a series of paired eye movements (Rayner et al., 2012, p.91) in which the first movement is a saccade (Matin, 1974) or superficial scanning of the sentence during which some words may be entirely skipped. Further, little or no comprehension of the scanned words occurs during a saccade (Matin, 1974, Rayner et al., 2012, p.91). The second movement is a fixation (Dearborn, 1906, Frazier & Rayner, 1982) or focusing of the eyes on the first word to the right of the last scanned character. After this word is read, the next saccade-fixation pair begins. Such saccade-fixation pairs take place in very short time intervals of about 20 – 35 milliseconds per saccade and between 150 and 500 milliseconds per fixation (Rayner et al., 2012, p.91) and go on until the end of the sentence is reached.

The comprehension stage may involve multiple sub-stages like parsing, understanding the syntax and context, interpreting the totality of the text, and reanalyzing the text if necessary (Rayner et al., 2012). Rather than going through so many sub-stages, however, people may rely on simple heuristics (Ferreira, Bailey, & Ferraro, 2002; Ferreira & Patson, 2007) to comprehend the text in fewer sub-stages with less cognitive effort. This is particularly likely when people have modest goals for the reading task (Rayner et al., 2012, p.242). For instance, they may not fixate on all the words of the sentence, understand their meanings, and then comprehend the sentence's syntactic and grammatical structure for its meaning. Instead, they may scan the sentence, fixate on just a few words and arrive at a meaning without ever decoding the syntax or grammar. Since the goals of interacting with tweets are also likely to be modest rather than highly significant, reading them may also be based on such heuristics. If words related to tweeting interests occur more often in the tweet, therefore, the likelihood of followers locating them as they scan it, fixating on them, and using them to comprehend the tweet should also increase. This in turn should increase retweets by those whose tweeting interests have a fit with what the tweet's topic is about.

The reading literature also finds that words on which readers fixate in the earlier parts of a sentence increase the likelihood of similar words being fixated on in the latter parts of the sentence (Engbert, Longtin, & Kliegl, 2002; Reichle, Rayner, & Pollatsek, 1999). Thus, earlier location of topic-related words in tweets is also likely to increase the likelihood of fixations on subsequently occurring topic-related words. This should increase the likelihood of the tweet's topic being recognized even if it is just scanned and thus increase retweets by those interested in the topic.

2.4. Categorization of topics and the role of topic-related words in reading

The literature on reading has also examined how people process and comprehend as they read (Ericsson & Kintsch, 1995; Kintsch, 1988; Potter, 1993). Griffiths et al. (2007) describe a model of comprehension in which readers make sense of words that they have read so far by tentatively associating a topic with those words. The topic associated with a word in turn is based on whether the word belongs to the group of semantically similar words that people categorize as a topic. An additional factor is the probability that they assign to whether the word is topic-related based on its frequency of occurrence in past encounters with that topic. Readers next use the tentatively identified topic to predict other words that they might encounter as they continue to read (Griffiths & Steyvers, 2003). Subsequently read words are then comprehended as adding additional meaning to the topic identified thus far unless those words appear to be part of a different topic in which case the inferred topic might itself be revised.

As an example of their model, Griffiths et al. (2007, p. 211) describe how the word “bank” may lead to the inference of a finance-related topic and the expectation of words like “federal” and “reserve” associated with that topic. If, however, the word “stream” is also encountered, the word “bank” may be used to conclude that the topic may be related to rivers leading to the expectation of words like “woods” and “field” rather than “federal” and “reserve”. The two key features of Griffiths et al. (2007) model are therefore that (1) readers start inferring the topic of a text as soon as they read a word and (2) the inferred topic has dynamic effects on readers' predictions of other topic-related words that are likely to occur and leads them to look for those words. Thus, in the case of a tweet, early occurrence of a topic-related word and the scanning for additional occurrences of other words related to the topic would lead to the conclusion that the tweet is about that topic if these words do occur. In light of Zhang et al.'s (2016) findings that people are more likely to retweet if tweets are on topics that they are interested in,

Table 1

Brands in the four categories.

| Brand | Followers | # of tweets (# of SPTs) | Brand | Followers | # of tweets (# of SPTs) |
|--------------------------|-----------|-------------------------|----------------------------|-----------|-------------------------|
| Automobiles | | | Airlines | | |
| Audi | 226,497 | 184 (2) | Southwest Airlines | 1,287,304 | 151 (13) |
| Ford | 123,293 | 241 (7) | Air Asia | 371,634 | 194 (33) |
| Chevrolet | 101,666 | 122 (8) | Virgin America | 321,093 | 100 (4) |
| Toyota | 84,766 | 361 (34) | Delta Air Lines | 316,447 | 274 (51) |
| Nissan | 71,307 | 309 (17) | British Airways | 213,176 | 82 (5) |
| VW | 65,603 | 181 (20) | Air New Zealand | 63,325 | 65 (10) |
| Harley-Davidson | 66,034 | 448 (10) | Hawaiian Airlines | 34,157 | 256 (12) |
| Jeep | 72,372 | 307 (16) | Lufthansa (USA) | 28,123 | 108 (6) |
| BMW | 57,312 | 53 (1) | Air Canada | 26,087 | 236 (108) |
| Porsche | 55,627 | 368 (5) | Food & Beverage | | |
| General-Motors | 53,897 | 253 (16) | Brand | Followers | # of tweets (# of SPTs) |
| Honda | 45,050 | 306 (31) | Pepsi | 632,473 | 163 (9) |
| Dodge | 38,382 | 294 (12) | Red Bull | 602,439 | 146 (3) |
| Chrysler | 31,814 | 306 (12) | Monster Energy | 226,473 | 443 (11) |
| Tesla | 30,829 | 396 (11) | Dunkin' Donuts | 140,484 | 113 (27) |
| Aston-Martin | 29,161 | 416 (6) | Domino's Pizza | 129,335 | 86 (8) |
| Cadillac | 28,374 | 297 (10) | Dr Pepper | 94,617 | 98 (13) |
| Mazda | 27,822 | 267 (14) | PepsiCo | 69,313 | 62 (0) |
| Hyundai | 26,762 | 151 (11) | Gatorade | 66,807 | 286 (15) |
| Dining | | | Mountain Dew | 65,782 | 84 (4) |
| Brand | Followers | # of tweets (# of SPTs) | Ben & Jerry's | 54,784 | 272 (20) |
| Subway | 467,792 | 196 (33) | Wheat Thins | 48,464 | 347 (13) |
| McDonald's | 442,340 | 338 (17) | Arizona Iced Tea | 39,724 | 270 (10) |
| Hard Rock Cafe | 156,176 | 261 (4) | Oreo | 40,082 | 129 (65) |
| Chick-fil-A | 137,209 | 137 (11) | Sierra Nevada | 36,016 | 203 (9) |
| KFC | 85,388 | 154 (48) | Kraft | 32,814 | 286 (11) |
| Chili's Grill & Bar | 58,135 | 366 (62) | Bacardi | 32,666 | 450 (50) |
| Papa John's Pizza | 50,742 | 279 (67) | Skittles | 31,116 | 467 (2) |
| Arby's | 48,850 | 96 (6) | | | |
| P.F. Chang's | 43,536 | 53 (3) | | | |
| The Cheesecake Factory | 46,142 | 95 (1) | | | |
| Dairy Queen | 33,935 | 302 (16) | | | |
| Hooters | 30,539 | 354 (19) | | | |
| Sonic Drive-In | 22,120 | 103 (5) | | | |
| Outback Steakhouse | 19,435 | 110 (29) | | | |
| California Pizza Kitchen | 16,933 | 98 (4) | | | |
| Popeyes Chicken | 12,561 | 352 (13) | | | |
| T.G.I. Friday's | 11,968 | 225 (3) | | | |

therefore, early and multiple occurrences of topic-related words in a tweet should increase retweets by those interested in the topic.

3. Data and variables

3.1. Data for phases 1 and 2

The total number of followers of a brand is a key metric in social media performance (Hoffman & Fodor, 2010). Investigating the retweets of tweets by brands with the most followers in their categories thus reduces brand-heterogeneity in social media performance and hence in the ability to compose tweets that gain retweets.⁴ We selected tweets by twenty brands with the most followers in each of two product (automotive and food & beverage) and two service (airlines and dining) categories for the investigations in phases 1 and 2.

For each brand, we collected the most recent 500 tweets using the Twitter Public API (Twitter, 2017a). We therefore include the same number of tweets from each brand. This prevents any of the twenty brands from playing a dominant role in our findings. Since the number of followers of brands may vary over time and brands with more (fewer) followers may get more (fewer) retweets (Araujo et al., 2015; Petrovic, Osborne, & Lavrenko, 2011; Suh, Hong, Piroli, & Chi, 2010), we also acquired data on the number of followers of each brand on the posting date of each of its collected tweets.⁵

⁴ As discussed subsequently, we also control for observed differences between brands in the number of followers as well as unobserved differences through brand random-effects.

⁵ This data is available from fanpagelist.com.

Table 2
Distribution of Retweets.

| | |
|--------------------|-------|
| Mean | 15.13 |
| Standard deviation | 66.67 |
| Minimum | 0 |
| Median | 5 |
| Maximum | 3131 |

3.2. Date setup – filtering tweets and brands

Tweets can either be “normal tweets” or replies (Twitter, 2017b). Brands broadcast normal tweets to all of their followers but send replies only to the followers who tweet a question or a complaint. We only retained normal tweets for our analysis and also excluded tweets that included an incentive for retweeting. Further, since we include brand random effects in our investigation, we dropped brands with fewer than 50 normal tweets to ensure that we can reliably update the mean and variance of each brand’s random-effect during the MCMC iterations. Our final sample consists of 14,148 tweets by 62 brands across the four product categories posted between February 4th, 2011 and April 26th, 2012. Table 1 lists the brands, the number of their followers on April 26, 2012, and the total of their normal tweets in our sample. We used a random sample of 11,318 tweets (80% of the total) for model calibration and the remaining 2830 tweets for validation testing.

3.3. Retweets

We recorded the number of retweets of each tweet in our final sample also using the Twitter Public API on April 1, 2013, which was about one year after collecting the tweets. Past empirical research on retweets has typically assumed that most retweets occur within a week of posting (e.g., Zaman, Fox, & Bradlow, 2014). Recording the number of retweets about one year or later from the posting date should therefore reduce the likelihood of undercounting retweets. Table 2 gives the distribution of retweets.⁶

3.4. Identifying Sales Promotional Tweets (SPTs)

We focus on sales promotional tweets that offer non-monetary incentives like free gifts rather than monetary incentives. This reduces the likelihood that the monetary values of the offered price reductions could affect the number of retweets. To identify such SPTs in our sample, we take the following approach. We first identify all unique stemmed words (Zhang et al., 2016) in the tweets. We then compute the frequency of occurrence of each stemmed word by itself or as part of a two or a three-word phrase and retain all words that occur at least 50 times and all phrases that occur at least 10 times. From the words that occur at least 50 times, we select those that are likely to be used in sales promotions that offer non-monetary incentives like free gifts rather than monetary incentives. There were seven such words: *chance*, *commercial*, *free*, *gift*, *giveaway*, *promo* and *win*. The word *sale* also occurred more than 50 times. Despite its association with price reductions, we include this word also in this category because it could also be used in sales promotions that do not offer price reductions.

Overall, therefore, we consider eight promotional words in our analysis. Our reliance on frequency of occurrence of these semantically-related words to define the topic is subjective but consistent with the Griffiths et al. (2007) model of how people categorize semantically similar words into topics.⁷

3.5. Presence of other topics in SPTs

SPTs are unlikely to only contain words related to sales promotions and can include additional words related to other topics as well. For instance, a brand’s tweet about a sales promotion organized to coincide with St. Patrick’s Day may include references to it and specific days and times of the promotion in addition to the promotional words. If people categorize *St. Patrick’s Day* with words related to cultural events, retweets of the SPT could also be affected by their (dis)interest in the topic. Ignoring the effects of words unrelated to promotions would therefore bias the estimated effects of words related to promotions on retweets of SPTs. We thus need to identify other topics in the SPTs and control for their effects on retweets in phases 1 and 2.

One approach to identifying additional topics in SPTs is to use topic modeling methods like Latent Dirichlet Allocation (Blei et al., 2003; Tirunillai & Tellis, 2014; Zhang et al., 2016). A disadvantage of LDA and other similar methods (Mimno, Wallach, Talley, Leenders, & McCallum, 2011; Wei & Croft, 2006) is that the set of words identified as a topic may “emphasize syntactic structure over semantic similarity” (Griffiths et al., 2007, p.212). The topics may therefore group together words because they co-occur even if they are not necessarily semantically similar (Griffiths et al., 2007).

Given the increasing use of LDA in academic marketing research, however, we investigated its ability to identify the top five and top ten categories of words from the 18,590 stemmed words in our sample of tweets. Consistent with concerns expressed in the literature, the words within these topics did not present a semantically or managerially coherent pattern (the identified topics are in Table 3).

⁶ Twitter does not include any modified versions of the original tweet or their retweets in the retweet count.

⁷ In phase 3 however we use topics identified within the tweets themselves by their authors. We discuss this in more detail shortly.

Table 3

Topics and words in topics with Latent Dirichlet allocation.

| 5 Topics – Top 10 words in each topic | |
|--|--|
| Topic 1 | see, love, just, thanks, good, best, car, bacardi, gt, great |
| Topic 2 | happy, today, get, happy, like, one, know, time, chicken, photo |
| Topic 3 | rt, win, w, chance, first, last, congrats, enter, game, amp |
| Topic 4 | time, check, see, can, watch, hooters, video, year, week, harleydavidson |
| Topic 5 | us, u, will, want, free, jeep, show, md, help, toyota |
| 10 Topics – Top 10 words in each topic | |
| Topic 1 | see, new, watch, gt, video, jeep, md, car, chrysler, ff |
| Topic 2 | happy, weekend, bacardi, today, porsche, tesla, via, will, s, tonight |
| Topic 3 | rt, love, thanks, like, good, chicken, just, ever, popeyeschicken, nice |
| Topic 4 | time, first, one, go, know, team, cadillac, going, way, top |
| Topic 5 | us, win, w, chance, enter, tweet, tomorrow, amp, live, join |
| Topic 6 | get, can, will, big, game, ready, help, just, back, fans |
| Topic 7 | check, now, awesome, airasia, rules, canada, see, code, official, details |
| Topic 8 | new, u, want, show, best, honda, year, toyota, nissan, gm |
| Topic 9 | day, great, today, week, harleydavidson, photo, ride, last, sure, facebook |
| Topic 10 | free, hooters, rt., favorite, pizza, skittles, vote, congrats, better, thank |

In light of the low semantic similarity of words within the categories identified by LDA, we took an alternative approach to group semantically similar non-promotional words of the SPTs into topics. Specifically, we defined four categories of words that are likely to be used in SPTs and selected words in those categories that occurred more than 50 times. The specific categories that we use are: (1) brand names (2) cultural events (e.g., St. Patrick's Day, Holiday, NCAA) (3) words related to time (because SPTs may include start and end days and times of the promotions) and (4) calls-to-action like 'click' and 'join'. In subsequent notation, these categories are labeled as *Brand*, *Events*, *Time*, and *CTA* respectively. Table 4 lists the promotional words and words in the four non-promotional categories.

Our approach to the categorization of words into categories above is in the spirit of Griffiths et al. (2007)'s model of how people categorize words that are semantically similar. We note, however, that we only use this approach to define topics in the context of phases 1 and 2. As mentioned previously, our objective in phase 1 is to investigate for prima facie evidence that the number and locations of topic-related words in tweets affect retweets. The objective of phase 2 is to investigate whether the share and locations of promotional words in these tweets *cause* retweets. Both phases are therefore used for evidence that either supports or refutes the role of the location of and space devoted to topic-related words in retweets. In phase 3 where we test for the reliability and generalizability of the prima facie evidence, however, we do not define topics based on the frequency of occurrence of words that we view as semantically similar. Instead, we define the topic of a tweet in this phase as the word following the “#” symbol. This is how Twitter defines a topic and one of the ways that readers may use to readily identify a tweet's topic.

3.6. Operationalization of the location of promotional and non-promotional words

For each of the five categories of words, we define variables, $Loc_{i, Prom}$, $Loc_{i, Brand}$, $Loc_{i, Events}$, $Loc_{i, Time}$, and $Loc_{i, CTA}$ which represent the location in number of characters from the start of the tweet where the first word from the Promotional-words, Brand, Events, Time, and CTA, categories respectively appears in the tweet. Previous findings in the literature (Malhotra et al., 2012) however indicate that the length of a tweet has a negative effect on retweets. This may be indicative that early occurrence of topics in longer tweets plays a larger role in retweets than in shorter tweets. A topic that occurs at the 20th character in a tweet of 140 characters for instance may be more effective in generating retweets than one that occurs at the 20th character of a tweet that is 40 characters long. Because tweets can vary in length, we therefore measure the location of the first occurrence of a topic-related word as the ratio of the number of characters from the start of the tweet where the word begins and the tweet's total length in characters. Thus, for instance, if the first word from the promotional-words category appears at the 35th character of the tweet and the tweet is 140 characters long, we set $Loc_{i, Prom}$ to 0.25. If no word from a category is present in the tweet, we set the location for that category to zero. This ensures that this category has no effect on the number of retweets.

Table 4

Words associated with promotional and non-promotional words.

| Topic | Words | Number of tweets with words |
|-----------------|---|-----------------------------|
| Promotions | free, chance, sale, commercial, giveaway, promo, win, gift | 1096 |
| Brand | Any of the 62 brands in the data | 3743 |
| Events | game, Event, holiday, NCAA, valentine, bowl, spring break, St Patrick's Day | 582 |
| Time | Saturday, Sunday, Monday, Tuesday, Wednesday, Thursday, Friday, morning, day, night, today, now, year, week, weekend, tomorrow, tonight | 4142 |
| Calls-to-action | click, watch, join, vote | 1557 |

3.7. Operationalization of the space devoted to promotional and non-promotional words in tweets

We operationalize the space devoted in a tweet to words from each of the five categories as the share of the total number of characters in the tweet occupied by characters in the words from the categories that appear in the tweet. Thus, for instance, the share of occurrence of any of the eight words for the promotional-words category (Table 4) is operationalized as:

$$Shr_{i, Prom} = \frac{\sum_{n=1}^8 I_{in}(Prom_n) \cdot Char(Prom_n)}{Length(i)} \times 100 \quad (1)$$

$I_{in}(Prom_n) = 1$ if tweet i includes sales-promotional word n , $Char(Prom_n)$ = total number of characters in the sales-promotional word n in the tweet, and $Length(i)$ = total number of characters in tweet i . $Shr_{i, Brand}$, $Shr_{i, Events}$, $Shr_{i, Time}$, and $Shr_{i, CTA}$ are analogous operationalizations for the *Brand*, *Events*, *Time* and *CTA* categories respectively.

3.8. Other variables

3.8.1. Symbols and acronyms

The “Twitter Lexicon” includes several symbols, acronyms and writing conventions (The Wall Street Journal, 2013). For instance, as mentioned previously, letters that follow ‘#’ up to the next blank space are one indication of the tweet’s topic to the readers. Similarly, the acronym ‘RT_if’ alerts the reader that the sender is requesting a retweet. Likewise, the presence of one or multiple ‘\$’ signs may signal a sales promotion. Tweets can also include multiple blank spaces that readers can complete. A tweet that includes multiple blank spaces which may have to be filled in before retweeting however can also take more time to read. Because some or all of these characteristics can affect the number of retweets (Malhotra et al., 2012), we define indicator variables $I_i(\#)$, $I_i(RT_{if})$, $I_i(\$)$, and $I_i(\text{blanks})$ (for any sequence of multiple blank spaces in the tweet). Additionally, because links to videos and photos increase retweets (Araujo et al., 2015), we also define an indicator variable $I_i(HTTP)$ for the presence of such links in the tweet.

3.8.2. Punctuations

Punctuations can also affect retweeting decisions (Araujo et al., 2015). For instance, ‘!’ can indicate that the tweet may have content of interest and ‘?’ may indicate that the recipient needs to think about the question being asked thus increasing the time to read and possibly reducing the number of retweets (Malhotra et al., 2012). We, therefore also define indicator variables $I_i(!)$ and $I_i(?)$. Additionally, because several tweets in our data had multiple exclamations, we define a variable $NEXCL_i$ for the number of exclamations in the tweet.⁸

3.8.3. Length

Longer tweets get fewer retweets (Malhotra et al., 2012). We therefore define a variable $Length(i)$ which is a standardized count of the number of characters in the tweet as an additional variable.⁹

3.9. Variables for posting time

The day on which a tweet is posted can affect its retweets (Araujo et al., 2015). Additionally, practitioners find that the time of posting can also affect retweets (Inc, 2013) – a finding confirmed in empirical investigations in the information systems literature (Petrovic et al., 2011). We therefore define indicators $I_i(\text{Saturday}) \dots I_i(\text{Thursday})$ for the posting day leaving Friday as the baseline and $I_i(0 : 00 H - 4 : 00 H) \dots I_i(16 : 00 H - 20 : 00 H)$ for each four hour block of Central Standard Time¹⁰ with 20:00H – 24:00H as the baseline. In addition, because retweets can also be affected by the number of days since posting (Zaman et al., 2014), we also define a variable, $Days_i$, for the number of days since a tweet was first posted.

3.9.1. Number of tweets

Given the little time spent on Twitter, a large number of tweets by a brand within a short time could lead people to ignore some tweets thus reducing their retweets. For each tweet of each brand in our sample therefore we also define a variable for the total number of tweets N_{jt} by brand ‘j’ on the day it was posted.

3.9.2. Number of followers

Brands with more followers get more retweets of their tweets (Zadeh & Sharda, 2014). We therefore also define a variable F_{jt} for the number of followers of brand j for each date on which it posted a tweet in our sample.¹¹

⁸ 6108 tweets in the sample had a single exclamation and 336 tweets had multiple exclamations.

⁹ Because the share variables include $Length(i)$ in the denominator, including it again in the model could lead to issues of collinearity. Because we use a standardized count of the number of characters as $Length(i)$, however, but use the actual length of the tweet in number of characters to compute the shares, collinearity should not be a concern. The correlations between $Length(i)$ and $Shr_{i, Prom}$, $Shr_{i, Brand}$, $Shr_{i, Time}$, $Shr_{i, CTA}$, and $Shr_{i, Events}$ were only 0.03, 0.07, 0.12, 0.04 and 0.14 respectively.

¹⁰ Although Twitter is used by consumers globally, we relied on posting time in the US because almost all of the brands in our study are American. All the times are converted to Central US time.

¹¹ We obtained this information from fanpage.com.

Table 5

Control variables in the data.

| Variables | Operationalization |
|--|--|
| Symbols and acronyms | Indicators for whether the tweet has: HTTP, RT, if, #, \$, sequence of multiple blank spaces |
| Punctuations | Indicators for whether the tweet has: ?, ! and number of ! |
| | Indicators for the day of the week of the tweet (Friday is the base) |
| Day and time of the tweet and elapsed time | Indicators for the six four-hour blocks of Central Standard Time (20:00–24:00 is the base) when the tweet was posted |
| | Number of days since the tweet was posted |
| Number of tweets | Total number of tweets posted by brand on the day of the tweet |
| Number of followers | Number of followers on Twitter on the day of the tweet |
| Following other brands | Number of others being followed by brand on the day of the tweet |
| | Indicator for brand's presence on Google+ |
| | Indicator for brand's presence on YouTube |
| | Indicator for brand's presence on Pinterest |
| Presence on other social networks | Indicator for brand's presence on Instagram |
| | Number of fans on Facebook on the day of the tweet |
| | Number of fans gained on Facebook on the day of the tweet |
| Brand resources | Presence/absence of brand on the Interbrand 100 list |
| Product categories | Indicators for the four product categories |

3.9.3. Following other brands

Several brands also participate in brand networks and follow other brands to increase their visibility in social media (Gensler, Völckner, Liu-Thompkins, & Wiertz, 2013). Because increased visibility of the brand could affect the number of retweets, we also define a variable $FOLL_{jt}$ for the number of other Twitter accounts followed by a brand on the posting date of each of its tweets in our sample.

3.9.4. Presence on other social networks

Retweets could also be affected by brands' presence on other major networks like Facebook. For instance, the brand may inform its fans on Facebook about a tweet that it posted and thus increase interest and retweets. To control for such effects, we define two variables – one for the number of fans $FANS_{jt}$ on the posting date of the tweet and another $FANGAIN_{jt}$ for the number of fans gained on that date on Facebook. We also define indicator variables for brands' presence on Google+ ($I(G_{+j})$), YouTube ($I(YT_j)$), Instagram ($I(INST_j)$), and Pinterest ($I(PIN_j)$)¹².

3.10. Brand resources

To control for unobserved differences between brands in the resources they are able to devote to social media and hence possibly in their abilities to compose tweets that gain retweets, we define an indicator variable, ($I(IB_j)$). This is an indicator for whether the brand is part of the list of the 100 largest brands on Interbrand rankings (Interbrand, 2012). As an additional control for differences between brands, we also include brand random effects.

3.11. Differences between categories

Because product categories vary in how much interest they attract from consumers (Berger & Schwartz, 2011), we also control for category differences with category indicator variables, $I(AUTO_k)$, $I(AIR_k)$, $I(F \& B_k)$, and $I(REST_k)$ corresponding to each category in our data. Table 5 lists all the variables.

3.12. Model free evidence

Tables 6 provides model-free evidence of the role of the shares and locations of promotional words in retweets. Comparing tweets where the average location of the first promotional word is earlier/later than the average location for SPTs shows a pattern. The mean number of retweets is higher for tweets with an earlier than average location of the first promotional word. The table also compares the mean number of retweets between tweets where the average share of promotional words is higher/lower than the average share of promotional words for all SPTs. The table indicates that the mean number of retweets is higher for SPTs with a higher than average share of promotional words. The table thus suggests a positive relationship between the early location of and higher space devoted to promotional words in SPTs and the number of retweets.

¹² The membership status on these networks did not change during the period of our data for any of the investigated brands.

Table 6

Effect of share and location of sales-promotional words in tweets that include them.

| Mean number of retweets for Promotional tweets with $Loc_j > mean$ | Mean number of retweets for Promotional tweets with $Loc_k < mean$ |
|--|--|
| 10.52 | 16.62 |
| Mean number of retweets for Promotional tweets with $Shr_j > mean$ | Mean number of retweets for Promotional tweets with $Shr_j < mean$ |
| 14.92 | 12.84 |

4. Model

The median number of retweets in our sample is 5 but about 30% of the tweets have 11 or more retweets. We therefore treat the number of retweets as over-dispersed count data and use a hierarchical Poisson Log-normal mixture to model them. The Log-normal mixture captures heterogeneity better than the Gamma mixture (Winkelmann, 2013). The total number of retweets, y_{ijkt} , of tweet i by brand j in category k on date t , is therefore modeled as

$$y_{ijkt} \sim \text{Poisson}(\lambda_{ijkt} * \epsilon_i) \\ \epsilon_i \sim \text{lognormal}(0, \tau_\epsilon) \quad (2)$$

We model λ_{ijkt} as a hierarchical function of the locations of the first word of each of the five categories of words, space devoted to the words related to each category in the tweets, and the other variables defined previously:

$$\log(\lambda_{ijkt}) = I_i(Prom) \cdot (\eta_{Prom} + \gamma_3 Shr_{i,Prom} + \theta_3 Loc_{i,Prom}) + I_i(Brand) \cdot (\eta_{Brand} + \gamma_1 Shr_{i,Brand} + \theta_1 Loc_{i,Brand}) \\ + I_i(Events) \cdot (\eta_{Events} + \gamma_2 Shr_{i,Event} + \theta_2 Loc_{i,Events}) + I_i(Time) \cdot (\eta_{Time} + \gamma_4 Shr_{i,Time} + \theta_4 Loc_{i,Time}) \\ + I_i(CTA) \cdot (\eta_{CTA} + \gamma_5 Shr_{i,CTA} + \theta_5 Loc_{i,CTA}) + \beta_X \bar{X}_i + (\delta_1 \cdot I(AUTO_k) + \delta_2 \cdot I(AIR_k) + \delta_3 \cdot I(F\&B_k) + \delta_4 \cdot I(REST_k)) + \gamma_j^t \quad (3)$$

- $I_i(Prom) = 1$ if tweet i has a promotional word; 0 otherwise; the other indicator variables $I_i(Brand)$, $I_i(Events)$, $I_i(Time)$ and $I_i(CTA)$ are analogously defined.
- η_{Prom} , η_{Brand} , η_{Events} , η_{Time} and η_{CTA} are the main-effects of the five word categories.
- $X_i = (I_i(\#), I_i(RT_{if}), I_i(), I_i(Blanks), I_i(HTTP), I_i(!), NEXCL_i, I_i(?), Length(i), I_i(Saturday) - I_i(Thursday), I_i(0:00H - 4:00H) - I_i(16:00H - 20:00H), Days_i)$
- $\delta_1 - \delta_4$ are the fixed-effects of the four product and service categories
- γ_j^t = effect of brand j on date t

Next, we model γ_j^t as the sum of a static component θ_j and a dynamic component that depends on the number of followers, followers gained, brands being followed, number of fans and fans gained by the brand on the day of the tweet:

$$\gamma_j^t = \theta_j + \gamma_1 F_{jt} + \gamma_2 FOLL_{jt} + \gamma_3 FANS_{jt} + \gamma_4 FANGAIN_{jt} + \gamma_5 N_{jt} \quad (4)$$

Finally, we model θ_j as a function of the brands' presence on the social networks Google+, YouTube, Instagram and Pinterest and the brands' resources:

$$\theta_j = \alpha_1 I(G_{+j}) + \alpha_2 I(YT_j) + \alpha_3 I(INST_j) + \alpha_4 I(PIN_j) + \alpha_5 I(IB_j) + \omega_j \quad (5)$$

ω_j is a mean zero brand random effect with unknown variance that captures the effects of unobserved brand attributes.

Table 7

Comparison of specifications.

| Specification of the Logit link | DIC | In-sample MSE | Out-of-sample MSE |
|---|--------|---------------|-------------------|
| Poisson Log-normal mixture (proposed specification) | 58,716 | 4888 | 1507 |
| Log linear regression | 72,564 | 4918 | 1499 |
| Poisson Log-normal mixture with only share and location effects | 58,792 | 5035 | 1523 |
| Poisson Log-normal mixture with share and location effects common across topics | 58,721 | 4891 | 1508 |

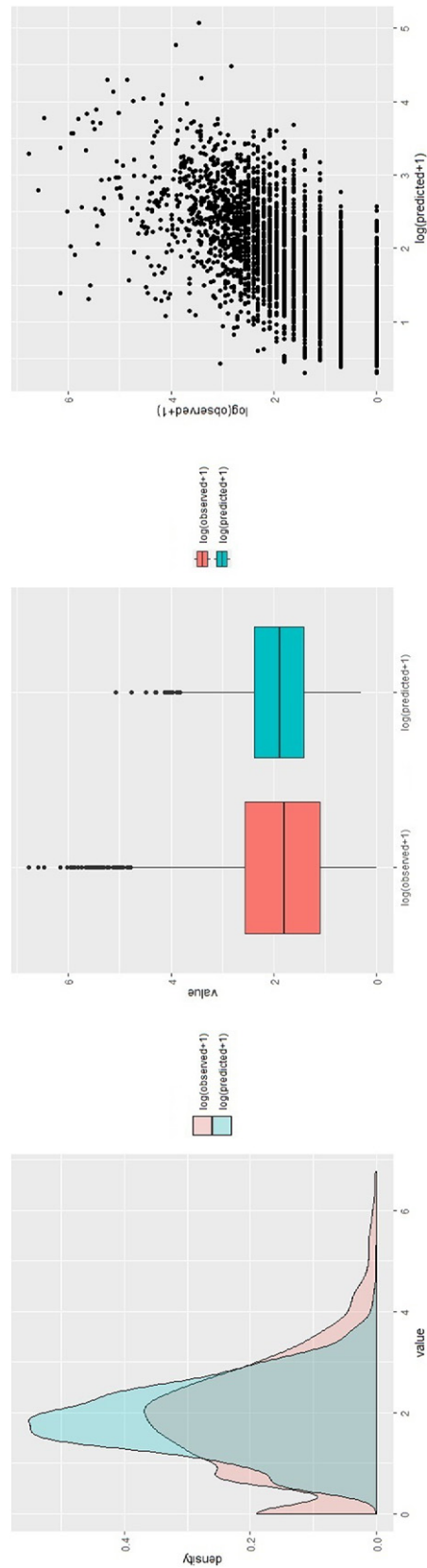


Fig. 2. Predictive Performance of the Model.

Table 8

Role of share and location of topic-related words.

| Topic | Mean | 2.50% | 97.50% |
|---|--------|---------|--------|
| Main effects of promotional and non-promotional words | | | |
| Promotions* | 0.521 | 0.313 | 0.753 |
| Brand* | 0.125 | 0.024 | 0.230 |
| Time* | 0.242 | 0.153 | 0.336 |
| Calls-to-action* | 0.301 | 0.082 | 0.529 |
| Events* | −0.264 | −0.514 | −0.006 |
| Share of promotional and non-promotional Words | | | |
| Promotions | 0.336 | −3.963 | 4.655 |
| Brand* | 1.644 | 0.677 | 2.582 |
| Time* | 1.182 | 0.459 | 1.880 |
| Calls-to-action* | −8.362 | −13.560 | −3.198 |
| Events* | 7.500 | 4.820 | 10.120 |
| Location of the First promotional and non-promotional words | | | |
| Promotions* | −0.642 | −0.948 | −0.358 |
| Brand | −0.130 | −0.287 | 0.020 |
| Time* | −0.291 | −0.446 | −0.140 |
| Calls-to-action | −0.050 | −0.276 | 0.178 |
| Events | −0.184 | −0.570 | 0.209 |

* 95% confidence interval does not include zero.

5. Phase 1: retweets of sales promotional tweets

We estimated the model using WinBUGS with a $N(0,1000)$ prior on all the parameters using the first 10,000 draws for burn-in and an additional 50,000 draws for inference. We also tested the model's specification by estimating three alternative specifications to assess the validity of the distributional and substantive assumptions:

1. A hierarchical Log-linear rather than a Poisson Lognormal regression to test whether the mixing distribution that we assumed is appropriate.
2. A specification that only includes main effects, and category-specific effects of the shares and locations of the five word categories, but no other variables. This tests whether the words that we selected for the five word categories do explain the number of retweets.
3. A specification that constrains the effects of the shares and locations of word categories to be common but the main effects to be specific to each of the five word categories. Comparing the fit of this specification with the model as specified in (2) – (5) tests whether the effects of share and location of promotional words on the number of retweets are different from those of other categories.

Table 7 compares the proposed specification with the three alternative specifications on both fit and in-sample and out-of-sample performance in predicting the number of retweets of tweets. It is clear from the comparison that the proposed specification does substantially better than the Log-linear regression on both fit and in-sample predictive performance and about the same on out-of-sample predictive performance. The fit of the Poisson Lognormal mixture with only share and location effects indicates that the shares and locations of words in each category are providing most of the explanation and predictive ability to the model. Finally, the table also indicates that the effects of shares and locations of words on retweets are category-specific because the proposed specification has the best fit and predictive performance.

To gain additional insights into the predictive performance of the model, we compare $\log(\text{observed retweets} + 1)$ of each tweet in the prediction sample with $\log(\text{predicted retweets} + 1)$ from three perspectives in Fig. 2. The kernel density plot at the top left of the figure suggests that the model over-predicts to some extent but is able to mimic the distribution of the observed retweets quite well. Next, the box-plot at the top right of the figure suggests that the means of observed and predicted retweets are quite close. Finally, the scatter plot in the bottom panel suggests that the model comes close to predicting large observed values of retweets although it over predicts small values. We next discuss the posterior summaries of model parameters¹³.

5.1. Main effects and effects of shares and locations of word categories (Table 8)

5.1.1. Main effects of word categories

The parameter for promotional words is the largest thus providing evidence that SPTs with promotional words are more likely to be retweeted than those with the other four categories. Among the main effects of the four non-promotional word categories, the parameter for *events* is also significant but negative suggesting that tweets by brands about events get fewer retweets than others. The effects of the other three categories of non-promotional words are positive and significant but vary in magnitude.

¹³ We also estimated a zero-inflated Negative Binomial (Winkelmann, 2008) specification to investigate whether we need to model any excessive zeros in the number of retweets as a separate process. There was no change in the magnitudes or significance levels of the parameters.

Table 9

Role of controls.

| Type of control | Variable | Mean | 2.5% | 97.5% |
|--|--|--------|--------|--------|
| Symbols and acronyms | Length of the Tweet* | −0.124 | −0.149 | −0.101 |
| | HTTP* | 0.150 | 0.096 | 0.200 |
| | RT_if* | 1.692 | 1.477 | 1.914 |
| | Hashtag (#)* | 0.068 | 0.023 | 0.115 |
| | Dollar Sign (\$)* | 0.638 | 0.524 | 0.756 |
| | Blank (___)* | −0.418 | −0.752 | −0.091 |
| Punctuations | Exclamation (!) | 0.032 | −0.074 | 0.128 |
| | Number of Exclamations* | −0.097 | −0.174 | −0.011 |
| | Question (?)* | −0.141 | −0.193 | −0.087 |
| | Saturday | 0.020 | −0.067 | 0.108 |
| Day and time of the tweet and elapsed time | Sunday* | 0.210 | 0.118 | 0.299 |
| | Monday* | 0.196 | 0.126 | 0.269 |
| | Tuesday* | 0.131 | 0.059 | 0.200 |
| | Wednesday* | 0.156 | 0.085 | 0.226 |
| | Thursday* | 0.102 | 0.030 | 0.172 |
| | Time of day - 0:4 | −0.040 | −0.224 | 0.143 |
| | Time of day - 4:8 | 0.114 | −0.034 | 0.257 |
| | Time of day - 8:12 | −0.002 | −0.112 | 0.100 |
| | Time of day - 12:16 | −0.066 | −0.170 | 0.034 |
| | Time of day - 16:20 | 0.004 | −0.112 | 0.112 |
| Number of tweets | Days Elapsed | −0.034 | −0.076 | 0.005 |
| | Total number of tweets posted by brand on the day of the tweet* | −0.211 | −0.232 | −0.190 |
| Number of followers | Number of followers on the day of the tweet* | 0.565 | 0.407 | 0.717 |
| Following other brands | Number of others being followed by brand on the day of the tweet | −0.211 | −0.232 | −0.190 |
| | Number of fans on Facebook on the day of the tweet | 0.160 | −0.020 | 0.325 |
| Presence on other social networks | Number of fans gained on Facebook on the day of the tweet * | 0.050 | 0.013 | 0.082 |
| | Indicator for brand's presence on Google+ | 0.113 | −0.177 | 0.571 |
| | Indicator for brand's presence on YouTube | −0.150 | −0.429 | 0.098 |
| | Indicator for brand's presence on Instagram | −0.106 | −0.483 | 0.208 |
| Brand resources | Indicator for brand's presence on Pinterest | −0.038 | −0.381 | 0.266 |
| | Presence/absence of brand on the Interbrand 100 list | −0.036 | −0.371 | 0.251 |
| | Automotive* | 1.955 | 1.596 | 2.512 |
| Product category differences | Food & beverage* | 1.323 | 0.933 | 1.733 |
| | Dining* | 1.354 | 1.022 | 1.761 |
| | Airline* | 1.098 | 0.617 | 1.585 |
| Precision of mixing Distribution | Precision of ϵ_i | 1.006 | 0.970 | 1.041 |

* 95% confidence interval does not include zero.

Table 10

Acronyms and Symbols Included in the Tweets in the Dataset.

| Acronym or symbol | Number of tweets with the acronym or symbol | Proportion of the sample of tweets with the acronym/symbol |
|-------------------|---|--|
| http | 8805 | 0.62 |
| RT_if | 145 | 0.01 |
| # | 5761 | 0.41 |
| \$ | 523 | 0.04 |
| ! | 6444 | 0.46 |
| ___ (Blank) | 72 | 0.01 |
| ? | 3121 | 0.22 |

The effect of *calls-to-action* is the largest followed by *time* and *brand* in that order. This pattern indicates that including calls to action along with timelines and references to the brand can increase retweets of SPTs.

5.1.2. Shares

The share of promotional words in SPTs does not have a significant effect indicating that the space devoted to them in the tweet does not affect the number of retweets. On the other hand, the shares of words related to *events*, *brand* and *time* have significant positive effects on retweets¹⁴. The effect of the share of event-related words however far exceeds that of words related to *brand* and *time*. This suggests that even a small increase in the share of event-related words in SPTs will overcome the negative main effect of this category of words. The parameter for *calls-to-action* however is negative and large indicating that brands should limit the use of such words. This is in contrast to Malhotra et al.'s (2012) recommendation that tweets that include words like “Looks” and “Today Only” that encourage action attract more retweets.

¹⁴ We re-estimated the model using an alternative operationalization of share as the number of topic-related words rather than the proportion of the tweet's length occupied by the words. The share of promotional words under this operationalization had a positive and significant effect on the number of retweets.

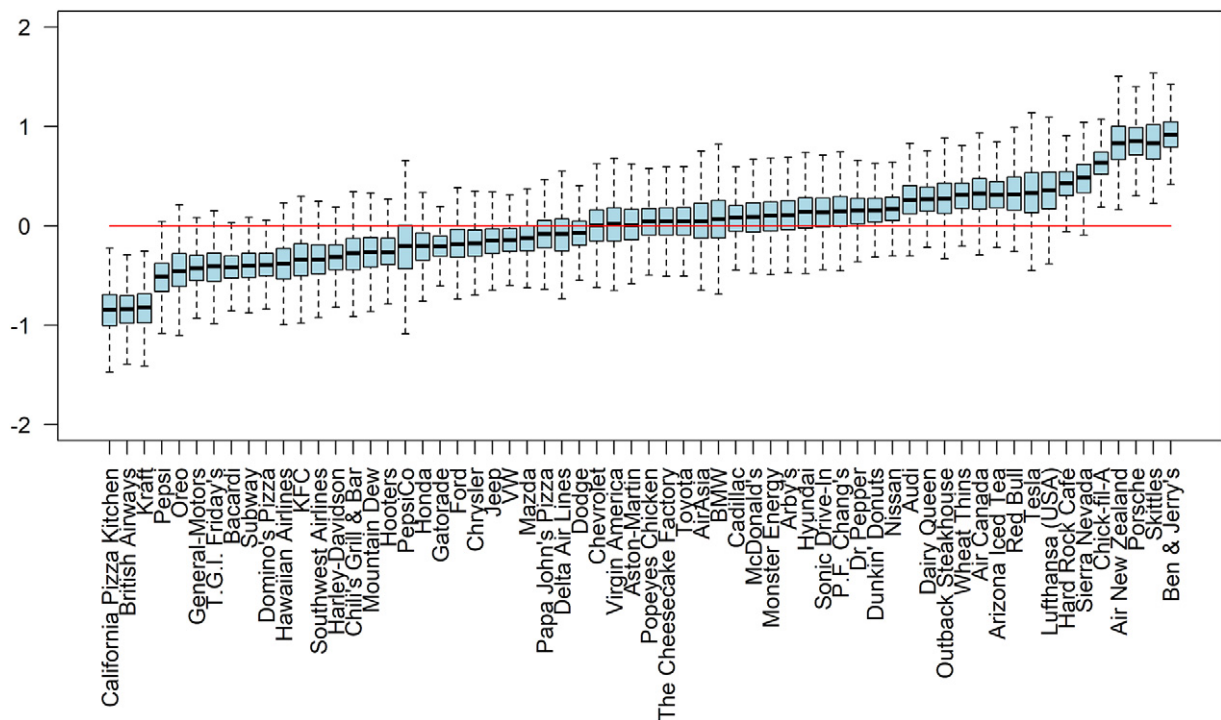


Fig. 3. Means of Random Effect Distributions of Brands.

5.1.3. Locations

The estimated parameter for the location of the first promotional word in an SPT is negative and significant¹⁵. Taken together with the non-significant parameter for the share of promotional words, this suggests that it is important for SPTs to be composed with promotional words early in the tweet. The only significant parameter for location from the four categories of non-promotional words is for time. Thus, in addition to the early occurrence of promotional words, SPTs should be written so that timelines also appear early in order to increase retweets.

5.2. Other Variables (Table 9):

5.2.1. Tweet characteristics

5.2.1.1. Length. The length of the tweet has a significant negative effect on the number of retweets indicating that an increase in the time needed to read the tweet decreases the number of retweets. This is consistent with past findings (Malhotra et al., 2012).

5.2.1.2. Symbols and acronyms. HTTP and RT_if have a positive relationship with the number of retweets although the parameter for RT_if is far higher and in fact is the largest of all the variables. Malhotra et al. (2012) report a similar positive relationship between the number of retweets and requests to retweet. A finding in our results which differs from one of their findings, however, is that # has a significant positive relationship with the number of retweets. This indicates that the ability to identify the tweet's topic easily increases retweets. Similarly, the presence of a \$ sign also increases retweets. On the other hand, blanks that allow the recipient to add text before retweeting reduce retweets.

5.2.1.3. Punctuations. A single exclamation has no significant effect but a question reduces retweets. Malhotra et al. (2012) report a similar relationship between question marks and retweets. Our estimates also indicate that the number of retweets is negatively related to the presence of multiple exclamations in the tweet.

5.2.1.4. Posting Time. Relative to the baseline of Friday, tweets posted on Sunday get the most retweets. The start of the week on Monday, however, leads to a reduction in retweets relative to Sunday with a further reduction on Tuesday while tweets in the middle of the week (Wednesday) get more retweets but fall on Thursday. This pattern is consistent with practitioner findings reported in the media (Fast Company, 2014). Our findings also suggest that posting time does not affect the number of retweets.

5.2.1.5. Number of tweets. Our results indicate that the number of retweets is negatively related to the total number of tweets posted by the brand on the same day. This suggests that posting several tweets on a day could reduce the number of retweets.

¹⁵ We re-estimated the model using the number of characters from the start of the tweet where the first promotional word appears as the operational measure of location rather than as the ratio of that number and the tweet's total length in characters. The location of the first promotional word was negative and significant under this operationalization as well.

5.2.1.6. Number of followers. The estimate for the number of followers suggests that the number of retweets increases with an increase in followers. This is not surprising however because the number of people interested in a tweet's topic is likely to increase as the number of followers increases.

5.2.1.7. Following other brands. The number of retweets however is not related to the number of others being followed by the brand. Thus, following individuals or other brands does not help brands to increase retweets of their tweets.

5.2.1.8. Maintaining a presence on other social networks. The estimates suggest that having fans on Facebook does not affect the number of retweets. On the other hand, the number of new fans gained on the day of the tweet is associated with more retweets. Presence on YouTube and Instagram, however, does not affect the number of retweets.

5.2.1.9. Brand resources. A brand's presence on the Interbrand index has no effect. Thus, an increase in the resources of brands is not associated with an increase in the number of retweets of their tweets.

5.2.1.10. Category effects. The estimated category effects arranged in descending order in the table are all significant indicating that tweets in all four categories receive at least some retweets. Tweets in the automotive category, however, are likely to get the most retweets followed by food & beverage, dining, and airlines.

5.2.1.11. Brand random effects. Because the MCMC iterations provide draws of the brand random effects, we also obtain the posterior distributions of the random effects for all the brands (Fig. 3). Box plots of the means indicate that brand-heterogeneity is being captured well by the model.

6. Phase 2: testing for causality

As mentioned previously, we use propensity score matching (Rosenbaum & Rubin, 1983) to assess whether there is a causal link between the shares and locations of words related to sales promotions and the number of retweets of SPTs. We next discuss our approach and results from this investigation.

We define high shares and earlier occurrence of promotional words as the treatments and use the following seven variables as the matching variables: tweet length, presence of HTTP, presence of hashtag (any hashtag), and the presence of words from the four non-promotional categories¹⁶. Matching on these characteristics from the same brand would generate treated and control samples of tweets matched on all parts of the tweet other than shares and locations of sales-promotional words. One constraint in our matching process is that we need to have a sufficiently large number of matched treatment and control sets of observations within each brand in order to estimate brand fixed-effects. We therefore consider the 20 brands with the most tweets in our data and use the following approach to generate a dataset of treated and control tweets for each brand:

1. Construct the empirical distribution of the share of promotional words in each SPT over the SPTs in the dataset.
2. Define a tweet as treated for the share of promotional words if the share of these words in the tweet is in the highest quartile of the empirical distribution from step (1). All other SPTs are the potential control sample of tweets.
3. Consider brand j , $j = 1 \dots 20$ in the selected set of 20 brands.
4. From the tweets by j in the dataset, using the nearest neighbor matching method, identify the treated and control tweets based on the matching variables. Discard the treatment tweets for which there are no matched tweets and combine the treated and control tweets into the sample for j .
5. Repeat steps 3 and 4 for all 20 brands and pool their samples into a single dataset.
6. Estimate the following negative binomial regression on the above dataset:

$$y_{ij} \sim \text{Poisson}(\lambda_{ij} * \epsilon_i), y_{ij} = \text{number of retweets of tweet } i \text{ by brand } j \\ \epsilon_i \sim \text{lognormal}(0, \tau_\epsilon) \quad (6)$$

$$\log(\lambda_{ij}) = \gamma_{T_{HighShare}} T_{HighShare_i} + \gamma_j \\ T_{HighShare_i} = \text{Indicator for whether } i \text{ is a treatment tweet} \\ \gamma_j = j\text{'s fixed effect} \quad (7)$$

7. Compute the empirical distribution of the location of the first promotional words (in number of characters) from the start of each tweet among the SPTs in the dataset
8. Define an SPT as treated for location if its first promotional word is located within the lowest quartile of the empirical distribution from step (7) (with this definition of treatment therefore the estimated effect of location on the number of retweets should be positive rather than negative). Use any of the other SPTs that match on the matching characteristics as controls.
9. Repeat steps 3–6, after replacing $T_{HighShare_i}$ with $T_{EarlyLocation_i}$ in step 6 where $T_{EarlyLocation_i}$ indicates whether i is a treatment tweet for location.

Table 11 presents the number of brands, the sample of tweets including treatment and control, and the estimated effects of high share and early location of promotional words from the above investigation. The effect of early location is significant at the 0.05 or better level

¹⁶ We could only use tweet length, HTTP, #, and other non-promotional words, since the other variables did not occur frequently and matching on them would result in a very small dataset to re-estimate the model (Table 10 gives the frequencies for all tweet characteristics other than length).

Table 11

Effects of the Share and Location of Topic-Related Words in Phase 2.

| | Number of brands | Number of tweets | Estimated coefficient and standard error |
|------------------------------|------------------|------------------|--|
| $\gamma_{Early_Location}^T$ | 8 | 95 | 0.410** (0.174) |
| $\gamma_{High_Share}^T$ | 8 | 54 | 0.337* (0.187) |

***, **, * significant at 1%, 5%, 10% respectively.

although the effect of share is only significant at the 0.10 level. Overall, however, the results are consistent with the findings from phase 1 and provide evidence that the locations of, and space devoted to, topic-related words affect the number of retweets of SPTs.

An implicit assumption that we make (as any propensity score matching based investigation would) in this analysis however is that unobserved conditions that affect the number of retweets are also matched for the treated and control tweets. If the unobserved conditions are different for the treated and control tweets, causality even if so indicated by a significant effect of the treatment (e.g., higher share or earlier location) would be questionable. For instance, the treated tweet could have been sent at a time when there is substantially more interest in promotions by the tweeting brand than when the control tweet was sent. Because this is not observed and hence not controlled by us, the larger number of retweets of the treated tweet would be incorrectly causally attributed to the higher share and/or location of promotion related words. Given the large corpus of tweets from which we carefully draw the treatment and matched samples, however, we believe that such events are unlikely to occur systematically for all treated tweets by all the brands that we investigate.

7. Phase 3: testing for endogeneity and generalizability

In this phase, we use a subset of a large corpus of about 1.77 million topic-author controlled tweets that result from a natural experiment on Twitter (Tan et al., 2014). Our goal is to rule out endogeneity and investigate the generalizability of the causal relationship of location and share of topic-related words identified in phase 2. Pairs of tweets in the topic-author-controlled dataset follow a pattern where one tweet is posted by an author at time t_1 and the same author posts a tweet subsequently at $t_2 > t_1$. Both tweets are on the same topic but are composed differently while using several words that might be common to both. The two tweets thus create a pair matched on the topic but differing in composition. They therefore create a natural experiment to study the effects of composition of tweets on retweets. We use only a subset of the 1.77 million tweets in this dataset to ensure a conservative test of whether (a) endogeneity is an alternative explanation for our findings and (b) our findings are generalizable beyond the topic of sales promotions. Specifically, we take the following approach to select the pairs of tweets.

First, to prevent any carryover effects of one tweet on the retweets of the other in the pair, we only select pairs in which $(t_1 + 24 \text{ hours}) \leq t_2 \leq (t_1 + 48 \text{ hours})$. The gap of at least 24 h should reduce the likelihood that the second tweet is seen as repetitive and ignored by the audience. Similarly, limiting the gap to 48 h should reduce the possibility that the second tweet gets more retweets because the first tweet's topic is forgotten and the second tweet is seen as new by the readers. Additionally, we only select a pair of tweets if it deals with exactly a single and same topic – as indicated by the occurrence of a hashtag followed by the same string of continuous characters in both tweets. Finally, to ensure that the tweets are composed mostly of the same words and only differ in their composition, we only consider pairs for which at least half of the words are common to both. Furthermore, in order to control for the effects if any of prolific tweeters and the potential effects of author expertise in writing tweets, we included the total number of tweets posted by each author as a control variable. These criteria resulted in a total of 18,406 tweets appearing in 9203 pairs and covering 4671 topics. We estimated the following negative binomial regression for the number of retweets of each tweet i in the sample:

$$\log(\lambda_i) = \beta_0 + \beta_1 \log(\text{Foll}_i) + \beta_2 \log(\text{NumTweets}_i) + \beta_3 \text{Length}_i + \beta_4 \text{Second}_i + \beta_5 \text{Hash.share}_i + \beta_6 \text{Hash.loc}_i + \beta_7 I_i(?) + \beta_8 I_i(!) + \beta_9 I_i(\text{Saturday}) + \beta_{10} I_i(\text{Sunday}) + \beta_{11} I_i(\text{Monday}) + \beta_{12} I_i(\text{Tuesday}) + \beta_{13} I_i(\text{Wednesday}) + \beta_{14} I_i(\text{Thursday})$$

Foll_i = number of tweets posted by i 's author on Twitter
 Length_i = i 's length in characters
 $\text{Second} = 1$ if i is posted second in its pair of tweets
 Hash.share_i = topic's share in the tweet measure as the length of the string of unbroken characters following # divided by the tweet's length in characters
 Hash.loc_i = ratio of the number of characters from the start of the tweet where the hashtag begins and the length of the tweet

(8)

Results of this regression presented in Table 12 show that the topic's share has a significant positive effect while its location has a significant negative effect on the number of retweets. These findings thus demonstrate that our results from phases 1 and 2 are not a result of endogeneity due to managerial expertise in composing SPTs that can attract more retweets. Additionally, they present evidence of the generalizability of our findings that earlier location of, and higher space devoted to, topic-related words in tweets increase retweets.

Table 12

Estimated parameters in phase 3.

| | Estimated coefficient | Standard error |
|------------------------------|-----------------------|----------------|
| (Intercept) | −5.027*** | 0.196 |
| $\log(\text{Folli})$ | 0.701*** | 0.008 |
| $\log(\text{NumTweet}_i)$ | −0.241*** | 0.016 |
| Length_i | 0.008*** | 0.001 |
| Second_i | −0.192*** | 0.032 |
| Hash. loc_i | −0.429*** | 0.048 |
| Hash. share_i | 3.396*** | 0.389 |
| Question mark | −0.215*** | 0.050 |
| Number of exclamation points | 0.041** | 0.019 |
| Weekday – Saturday | 0.074 | 0.061 |
| Weekday – Sunday | 0.212*** | 0.063 |
| Weekday – Monday | 0.155*** | 0.059 |
| Weekday – Tuesday | 0.208*** | 0.057 |
| Weekday – Wednesday | 0.124** | 0.056 |
| Weekday – Thursday | 0.083 | 0.058 |
| Time of day – 4:8 | −0.095 | 0.075 |
| Time of day – 8:12 | 0.317*** | 0.065 |
| Time of day – 12:16 | 0.099* | 0.054 |
| Time of day – 16:20 | −0.011 | 0.056 |
| Time of day – 20:24 | −0.037 | 0.060 |

***, **, * significant at 1%, 5%, 10% respectively.

8. Summary, managerial implications, and limitations

8.1. Summary

Retweets by followers can broadcast brands' tweets to large audiences. Findings in the literature however suggest that followers are more likely to retweet only if they recognize that the tweet's topic fits their tweeting interests. This is a challenge for brands since people spend only a few seconds reading a tweet. Followers may therefore read tweets cursorily, not recognize that the topic fits their tweeting interests, and not retweet. Brands therefore need insights into composing tweets to facilitate recognition of their topics even when they are just scanned. This is the issue that we address in this research. Specifically, we investigate whether locating topic-related words earlier in the tweet and increasing the space devoted to them in the tweet increases retweets.

Results from our empirical investigation of sales promotional tweets by brands suggest that early location of topic-related words increases retweets. This finding is confirmed in a subsequent investigation using propensity score matching methods. An additional investigation using topic-author-controlled pairs of tweets composed by individuals rather than brands demonstrates that earlier location of and higher space devoted to topic-related words in tweets increase retweets. Importantly, this analysis also provides evidence of the generalizability of our findings to tweets in general and not just those by brands on specific topics.

8.2. Managerial implications

Malhotra et al. (2012) cite the following benefits of retweets to make the case for brands to compose tweets so as to increase their “retweetability” (p.65):

1. Even a few additional retweets can increase a tweet's reach by the thousands.
2. Retweets add credibility to the tweet thus increasing its effectiveness by making it more likely to be read by those who are familiar with the retweeter.
3. Those who are not following the brand as yet may begin following the brand thus further increasing its network *and, hence, the reach of its tweets*¹⁷.

Malhotra et al. (2012), however, do not provide empirical evidence for the above reasons. To demonstrate the managerial value of our findings, we therefore investigated if retweets do gain new followers. Specifically, we considered the 19 brands in the automobile and 18 brands in the dining categories that we retained for our investigation and monitored their tweets, the retweets of those tweets by followers, and incremental changes in the number of followers of the brands following the tweets, between 5/3/2018 and 6/4/2018. The specific steps that we followed for collecting the data were as follows:

1. Record the number of followers $F_{i, t_{0i}}$ of brand i at t_{0i} when we start collecting the data on May 3rd 2018
2. Record the time t_{1i} of the first tweet after t_{0i} by brand i
3. Record the number of followers $F_{i, t_{1i}}$ of brand i at t_{1i}
4. Record the time t_{2i} of the tweet after the one at t_{1i}

¹⁷ The phrase in italics is our addition to Malhotra et al.'s (2012) reason.

5. Record the number of followers of the brand $F_{i, t_{2i}}$ of brand i at t_{2i}
6. Record the number of retweets, $rt_{t_{1i}, t_{2i}}$ during the period $[t_{1i}, t_{2i}]$ of the tweets by brand i at t_{1i}
7. Compute $G_{t_{1i}, t_{2i}} = F_{i, t_{2i}} - F_{i, t_{1i}}$ the increase in the number of followers of i between t_{2i} and t_{1i}
8. Repeat steps (2) – (7) for every subsequent consecutive pair of tweets at times $t_{r, i}$ and $t_{(r+1), i}$, $r = 2$, until the end of the monitoring period
9. Repeat steps (1) – (8) for all brands $i = 1 \dots 37$.

At the end of the monitoring period, therefore, we had a series of pairs of observations $(G_{t_{r, i}, t_{r+1, i}}, rt_{t_{r, i}, t_{r+1, i}})$, $r = 1 \dots N_{i-1}$, where N_i is the total number of tweets observed during the monitoring period for brand i . We were therefore able to investigate the relationship between the increases in the number of followers of each brand due to retweets. Specifically, we investigated the following log-linear as well as negative binomial regression:

| | |
|----|---|
| M1 | $\log(G_{t_{r, i}, t_{r+1, i}}) = \beta_0 + \beta_i + \gamma_1 \log(rt_{t_{r, i}, t_{r+1, i}}) + \gamma_2(t_{r+1, i} - t_{r, i})$ |
| M2 | $G_{t_{r, i}, t_{r+1, i}} \sim \text{Negbin}(\mu_{t_{r, i}, t_{r+1, i}}, \alpha_{t_{r, i}, t_{r+1, i}})$ $\log(\mu_{t_{r, i}, t_{r+1, i}}) = \beta_0 + \beta_i + \gamma_1 \log(rt_{t_{r, i}, t_{r+1, i}}) + \gamma_2(t_{r+1, i} - t_{r, i})$ |

In the first model, M1, the parameters β_i includes 36 dummy variables representing brand fixed effects, which control for additional sources of follower gain such as brand advertising. Additionally, because the frequency of tweets could affect follower gains, we also investigate both linear and non-linear effects of frequency using the time between tweets. Specifically, we include $t_{r+1, i} - t_{r, i}$ as well as its square in the model and thus estimate one additional version of both M1 and M2 that includes the non-linear effects of time.

We present the estimated effects of retweets on gains in followers by the brands in Table 13. The results in the table suggest that retweets of brands' tweets are positively associated with follower gains for all the brands. Because an increase in the number of followers can lead to greater interest in the brands and higher sales (Kumar et al., 2013, Gong et al., 2017, Petrova et al., 2017, Twitter, 2018, The Wall Street Journal, 2017), this analysis indirectly shows that an increase in retweets can also lead to increased sales. Our results demonstrating that retweets of brands' tweets increase their followers thus provide evidence of the managerial value of composing tweets to increase retweets.

8.3. Managerial application – composing tweets to increase retweets

Our research highlights the benefits of greater space devoted to and earlier location of topic-related words in tweets. In order to compose tweets that can attract retweets, therefore, social media managers should start by identifying topic-related words for the topics of their tweets. We suggest three approaches that brands can use to develop lists of topic-related words. One, they should monitor retweets of their tweets on the topic and identify the tweets that performed well on retweets. For instance, these could be tweets that receive more than the average number of retweets over all tweets on the topic. They should then compile a list of topic-related words in those tweets. Two, they should complement the lists of topic-related words compiled from approach one by monitoring tweets on the same topic by competitors. Topic-related words in competitors' tweets that receive more than the average number of retweets on all tweets on the topic by them should then be added to the internally formed lists. Third, they should run experiments by using alternative topic-related words to compose tweets on the same topic and select words from tweets that perform better on retweets. This approach is thus similar to the topic-author controlled experiments illustrated in Tan et al. (2014).

Once the lists of topic-related words are identified and continuously updated, we recommend the following approach for how social media managers can compose tweets on a topic to increase retweets¹⁸:

1. Place several topic-related words as close to the start as feasible.
2. Add a hashtag in front of each of the topic-related words to draw attention to the topic
3. Compose the rest of the tweet by filling it with relevant symbols and words as needed based on the following priority:
 - a. RT_IF and HTTP: The significant positive effects of RT_IF and HTTP imply that they should also be included in the rest of the tweet.
 - b. Other words, punctuations and symbols relevant to the message. For instance, a tweet about a new product may need to include "New", the name of the new product, and when and where the product would be available.
 - c. Words, punctuations and symbols necessary to complete the composition. For instance, verbs, pronouns, and periods have to be included as needed.

The overall approach therefore is to begin the composition of the tweet by placing words that can increase retweets in locations that can also help increase retweets and then completing the rest of the tweet with other words and symbols as needed. It is thus an alternative to composing and editing tweets with the primary goal of not exceeding the limit on the number of characters (Twitter, 2017c) and relying on including attention-grabbing words at the start (Malhotra et al., 2012) to increase retweets.

¹⁸ We thank the Senior Editor and an anonymous reviewer for encouraging us to develop this discussion.

Table 13

Effects of Retweets on Gains in Followers.

| | M1 | M1 + non-linear term | M2 | M2 + non-linear term |
|----------------------------------|------------|----------------------|-------------|----------------------|
| (Intercept) | 7.422*** | 7.438*** | 7.364*** | 7.384*** |
| $\log(rt_{t_r, i} - t_{r+1, i})$ | 0.034*** | 0.040*** | 0.045*** | 0.051*** |
| | | Retweets | | |
| | | Time between tweets | | |
| $(t_{r+1, i} - t_{r, i})$ | 2.09E-05** | −3.39E-05* | 2.10E-05*** | −3.95E-05*** |
| $(t_{r+1, i} - t_{r, i})^2$ | | 5.38E-09*** | | 5.56E-09*** |
| | | Brand fixed effects | | |
| Astonmartin ^a | 0.085 | 0.049 | 0.106*** | 0.064* |
| Audi | −0.017 | −0.056 | 0.199*** | 0.155*** |
| BMW | 0.127** | 0.087 | 0.149*** | 0.103*** |
| Cadillac | 0.112 | 0.076 | 0.129** | 0.087* |
| calpizzakitchen | 0.105 | 0.080 | 0.142** | 0.114** |
| Cheesecake | 0.063 | 0.030 | 0.087* | 0.051 |
| Chevrolet | 0.129 | 0.089 | 0.147*** | 0.101* |
| ChickfilA | 0.068 | 0.114 | 0.074 | 0.127** |
| Chilis | 0.088 | 0.054 | 0.110** | 0.072 |
| Chrysler | 0.065 | 0.026 | 0.078* | 0.032 |
| DairyQueen | 0.089 | 0.048 | 0.096* | 0.050 |
| Dodge | 0.049 | 0.001 | 0.045 | −0.010 |
| GM | 0.102 | 0.113* | 0.132*** | 0.144*** |
| HardRock | 0.104 | 0.075 | 0.139** | 0.107* |
| harleydavidson | 0.069 | 0.052 | 0.094*** | 0.075** |
| Honda | 0.099 | 0.062 | 0.116*** | 0.075* |
| Hooters | 0.087 | 0.079 | 0.121*** | 0.111*** |
| Hyundai | 0.154 | 0.126 | 0.187*** | 0.155*** |
| Jeep | 0.060 | 0.011 | 0.056 | 0.001 |
| MazdaUSA | 0.031 | 0.054 | 0.049 | 0.073* |
| McDonalds | −0.118 | −0.164** | 0.046 | −0.006 |
| Nissan | 0.022 | 0.010 | 0.038 | 0.024 |
| Outback | 0.108 | 0.076 | 0.137** | 0.101* |
| PapaJohns | 0.061 | 0.018 | 0.066 | 0.017 |
| PfChangs | 0.128* | 0.104 | 0.172*** | 0.144*** |
| Pizzahut | 0.038 | 0.046 | 0.063 | 0.071* |
| PopeyesChicken | 0.056 | 0.020 | 0.071 | 0.030 |
| Porsche | 0.126 | 0.073 | 0.227*** | 0.162*** |
| Sonicdrivein | 0.040 | −0.005 | 0.041 | −0.009 |
| SUBWAY | −0.085 | −0.129 | −0.047 | −0.097** |
| Tacobell | −0.053 | −0.035 | −0.028 | −0.007 |
| Tesla | 0.621 | 0.570*** | 0.716*** | 0.655*** |
| TGIFridays | 0.094 | 0.115* | 0.132*** | 0.154*** |
| Toyota | 0.122 | 0.089 | 0.147*** | 0.110*** |
| VW | 0.124 | 0.096 | 0.156*** | 0.123*** |
| Wendys | 0.541 | 0.487 | 0.720*** | 0.652*** |

***, **, * significant at 1%, 5%, 10% respectively.

^a Arby's is the base.

8.4. Contributions and limitations

As discussed previously, there have been other investigations of retweets in the literature. For instance, [Malhotra et al. \(2012\)](#) develop recommendations for how to increase retweets. Their findings however are based on a descriptive analysis of a much smaller dataset. Specifically, they analyze 1150 tweets by 47 brands (an average of less than 25 tweets per brand) while we rely on 14,148 tweets by 62 brands (an average of over 228 tweets per brand). [Zhang et al., 2016](#) also investigate retweets but focus on how the fit between a tweet's topic and the audience's retweeting interests plays a role. Our research thus adds to this literature by investigating a large dataset of tweets with a focus on how they should be composed in order to increase retweets. Specifically, the primary contribution of our research is in demonstrating that the locations of and space devoted to topic-related words in tweets affect retweets. Additionally, we use our findings to develop an approach for how social media managers of brands can compose tweets to increase retweets.

Our research however has some limitations. One question that we do not investigate is that of whether the effects of the locations of and space devoted to topic-related words on retweets vary with category characteristics. It is possible, for instance, that topic-related words play a bigger role in some categories than others. For instance, words related to promotions may play a larger role in the retweets of more expensive categories where promotions may be less frequent than in other categories. Similarly, the effects could also differ across brands based on how active they are on other social media. For instance, brands that post

a large number of messages on Facebook may not generate as many retweets if several of their followers on Twitter are also fans on Facebook.

We also do not investigate if the role of the locations of and space devoted to topic-related words in retweets depends on when the tweet is posted. For instance, it is possible that people are more interested in interacting with tweets during the evenings when they have more time for social interactions. If this is the case, the role of topic-related words may be stronger in the mornings than during evenings. Thus, earlier locations of topic-related words in the tweet might have a more positive effect during the day when people are likely to be busier than during the evenings and weekends.

Another limitation of our investigation is that we cannot fully account for cross-platform effects. For instance, tweets may stimulate discussions on other social media platforms like Facebook and Instagram which could in turn affect retweets. Such feedback effects could affect the magnitudes of the effects of the locations of and space devoted to topic-related words in tweets. This is therefore another issue that needs further investigation.

8.5. Future research directions

Our findings on how a larger share and earlier location of topic-related words highlight the importance of attracting attention to tweets' topics to increase retweets. They also raise several avenues for additional research on the composition of tweets to increase dissemination through retweets. One issue that we do not explore in this research is that of prioritization when a tweet addresses multiple topics. For instance, a recent tweet by the restaurant chain McDonald's on December 10, 2018 (Fig. 4) includes two distinct topics: the description of a sandwich and the introduction of the sandwich as a new product. Such multi-topic tweets are more likely because tweets can include up to 280 characters as opposed to the 140 character limit previously. Managers will therefore need insights into which topic(s) should be the leaders, i.e., closer to the start of the tweet, and which one(s) should be the follower(s) to increase retweets.

A second promising avenue relates to the sentiments in the topic-related words that are located closer to the start of the tweet. The tweet from McDonald's in the above example (Fig. 4) includes the phrase "seasoned, savory and satisfying" at the start. While



Fig. 4. A tweet by McDonald's on December 10, 2018.

all three words are related to eating the sandwich, each is associated with a different aspect of the experience. Which aspect is more likely to help the audience recognize the topic(s) faster and increase reading and retweeting is not clear. This is however likely to be related to the relative preferences of the brand's followers among these three aspects. An investigation, therefore, of the implications of followers' preferences to the composition of tweets and the shares and locations of topic-related words can help social media managers increase retweets.

Tweets can also include photos or videos. Another direction for managerially useful research therefore involves whether and how interactions between tweets' visual aspects and the location and share of topic-related words affect retweets. Returning to the McDonald's example in Fig. 4, how the sandwich is depicted is also likely to affect followers' interest in retweeting even after recognizing the tweet's topic. It is therefore important for scholars to draw on the increasing body of literature on visual social media and examine how visual features like colors (Jalali & Papatla, 2016) and visual complexity (Pieters, Wedel, & Batra, 2010) of photos included in tweets affect the roles of the shares and locations of topic-related words in retweets.

Given our focus on how a tweet's composition influences the total number of retweets, we also do not explore the role of network structures in retweeting. Research on the role of network structures in message propagation (Katona, Zubcsek, & Sarvary, 2011; Zubcsek, Chowdhury, & Katona, 2014) however suggest that the number of people that a brand's follower is connected to and the density of those connections (Zubcsek et al., 2014) can also affect the propagation of tweets through retweets. Investigations of the interplay between network structures and the composition of tweets in retweets gained by tweets is therefore another direction to build on our work.

Related to the above issue is an additional promising avenue for research that links messaging by brands through Twitter with their paid efforts for the same messages through advertising. Specifically, Zubcsek and Sarvary (2011) focus on another aspect of message transmission in social networks. Specifically, their interest is in investigating whether firms' that can communicate with social networks would be taking sub-optimal decisions on their paid advertising efforts if they ignore the word of mouth about those messages in the network. Their analytical results confirm that this will be the case. These findings suggest a promising direction for research. Specifically, an empirical investigation of whether firms that are able to increase retweets based on our findings, but ignore the effects of those retweets on message propagation, risk spending unnecessarily more on paid advertising to communicate the tweets' messages through paid media would be managerially useful.

Finally, another promising avenue for future research is the development of automated methods to generate tweets with the shares and locations of topic-related words designed so as to increase retweetability (Malhotra et al., 2012). This can substantially increase brands' ability to reduce the time and monetary investments needed to generate tweets that would be retweeted more. Automated methods being used to generate newspaper headlines that attract attention – for instance those that are in use at *The New York Times* (*The Wall Street Journal*, 2018) – could be adapted to compose tweets along these lines.

References

- Advertising Week (2014a). *Arby's slayed the Grammys with this tweet about Pharrell Williams' hat*. January, 27.
- Advertising Week (2014b). *Arby's dishes on awesome Pharrell Williams tweet*. (March 7).
- Advertising Week (2014c). *This is how much time we spend on social networks every day*. November 18th.
- Araujo, T., Neijens, P., & Vliegenthart, R. (2015 Sep 1). What motivates consumers to re-tweet brand content? *Journal of Advertising Research*, 55(3), 284–295.
- Barnes, N. G., Lescault, A. M., & Homes, G. (2015). The 2015 fortune 500 and social media: Instagram gains, blogs lose. *University of Massachusetts Dartmouth Center for Marketing Research*.
- Berger, J., & Schwartz, E. M. (2011). What drives immediate and ongoing word of mouth? *Journal of Marketing Research*, 48(5), 869–880.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Chae, I., Stephen, A. T., Bart, Y., & Yao, D. (2016). Spillover effects in seeded word-of-mouth marketing campaigns. *Marketing Science*, 36(1), 89–104.
- Company, F. (2014). *Twitter 101: How to get more clicks, retweets, and reach*. May, 13.
- Dearborn, W. (1906). *The Psychology of Reading*. Columbia University contributions to philosophy and psychology. New York: Science Press, 1906.
- Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in Reading based on spatially distributed lexical processing. *Vision Research*, 42(5), 621–636.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102, 211–245.
- Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11(1), 11–15.
- Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass*, 1(1–2), 71–83.
- Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14, 178–210.
- Gensler, S., Völckner, F., Liu-Thompkins, Y., & Wiertz, C. (2013). Managing brands in the social media environment. *Journal of Interactive Marketing*, 27(4), 242–256.
- Gong, S., Zhang, J., Zhao, P., & Jiang, X. (2017). Tweeting as a marketing tool: A field experiment in the TV industry. *Journal of Marketing Research*, 54(6), 833–850.
- Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. *Advances in Neural Information Processing Systems*, 11–18.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, 114(2), 211.
- Hinz, O., Skiera, B., Barrot, C., & Becker, J. U. (2011). Seeding strategies for viral marketing: An empirical comparison. *Journal of Marketing*, 75(6), 55–71.
- Hoffman, D. L., & Fodor, M. (2010). Can you measure the ROI of your social media marketing. *Sloan Management Review*, 52(1), 41–49.
- Inc (2013). "3 secrets of the Most shareable tweets," September 9.
- Inhoff, A. W., & Weger, U. W. (2005). Memory for word location during Reading: Eye movements to previously read words are spatially selective but not precise. *Memory and Cognition*, 33, 447–461.
- Interbrand (2012) (<http://www.bestglobalbrands.com/previous-years/2012>).
- Irwin, D. E. (1992). Visual memory within and across fixation. In K. Rayner (Ed.), *Eye movements and visual cognition: Scene perception and Reading* (pp. 146–165). New York: Springer-Verlag.
- Jalali, N. Y., & Papatla, P. (2016). The palette that stands out: Color compositions of online curated visual UGC that attracts higher consumer interaction. *Quantitative Marketing and Economics*, 14(4), 353–384.
- Kambe, G., Rayner, K., & Duffy, S. A. (2001). Global context effects on processing lexically ambiguous words: Evidence from eye fixations. *Memory & Cognition*, 29, 363–372.
- Katona, Z., Zubcsek, P. P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research*, 48(3), 425–443.

- Rayner, K., Alexander, P., Jane, A., & Charles, C., Jr. (2012). *Psychology of Reading*. Psychology Press.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182.
- Kumar, V., Bhaskaran, V., Mirchandani, R., & Shah, M. (2013). Creating a measurable social media marketing strategy: Increasing the value and ROI of intangibles and tangibles for hokey pokey. *Marketing Science*, 32(2), 194–212.
- Lovett, M. J., Peres, R., & Shachar, R. (2013). On brands and word of mouth. *Journal of Marketing Research*, 50(4), 427–444.
- Malhotra, A., Malhotra, C. K., & See, A. (2012). How to get your messages retweeted. *MIT Sloan Management Review*, 53(2), 61–66.
- Martin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin*, 81, 899–917.
- Mimno, David, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of The Conference On Empirical Methods In Natural Language Processing*, 262–272, Association for Computational Linguistics.
- Petrova, Maria, Sen, Ananya, and Yildirim, Pinar, Social Media and Political Donations, "New Technology and Incumbency Advantage in the United States (September 8, 2016)," Available at SSRN: <https://ssrn.com/abstract=2836323>
- Petrovic, S., Osborne, M., & Lavrenko, V. (2011). RT to win! Prediction message popularity in twitter. In *AAAI Conference on Weblogs and Social Media, AAAI. Spain: Barcelona*.
- Pieters, R., Wedel, M., & Batra, R. (2010). The stopping power of advertising: Measures and effects of visual complexity. *Journal of Marketing*, 74(5), 48–60.
- Potter, M. C. (1993). Very short term conceptual memory. *Memory & Cognition*, 21, 156–161.
- Rayner, K. (1998). Eye movements in Reading and information processing: 20 years of research. *Psychological Bulletin*, 124, 372–422.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in Reading: Accounting for initial fixation locations and Refixations within the EZ reader model. *Vision Research*, 39(26), 4403–4411.
- Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41–55.
- Stephen, A. T., & Galak, J. (2012). The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of Marketing Research*, 49(5), 624–639.
- Stephen, A. T., & Galak, J. (2012). The effects of traditional and social earned media on sales: A study of a microlending marketplace. *Journal of marketing research*, 49(5), 624–639.
- Stephen, A. T., & Lehmann, D. R. (2016). How word-of-mouth transmission encouragement affects Consumers' transmission decisions, receiver selection, and diffusion speed. *International Journal of Research in Marketing*, 33(4), 755–766.
- Suh, B., Hong, L., Piroli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. in *Social Computing (socialcom), IEEE Second International Conference*, 177–184.
- Tan, C., Lee, L., & Pang, B. (2014). The effect of wording on message propagation: topic- and author-controlled natural experiments on twitter. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) 2014*.
- The Wall Street Journal (2013). How Twitter's hashtag came to be. *October*, 3.
- The Wall Street Journal (2017). "You want Snark with those fries? No one is safe from Wendy's tweets: More than Most companies, the no. 3 burger chain has taken to twitter to taunt rivals and roast customers," July 25.
- The Wall Street Journal (2018). New York times adapts data science tools for advertisers. *February*, 15.
- Tirunillai, S., & Tellis, G. J. (2014). Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent Dirichlet allocation. *Journal of Marketing Research*, 51(4), 463–479.
- Townsend, D. J., & Bever, T. G. (2001). *Sentence comprehension: The integration of habits and rules*. Vol. 1950. Cambridge, MA: MIT Press.
- Twitter (2017a). downloaded from <https://support.twitter.com/articles/119138#>.
- Twitter (2017b). downloaded from <https://support.twitter.com/groups/50-welcome-to-twitter/topics/203-faqs/articles/13920-new-user-faqs>.
- Twitter (2017c). "Tweeting Made Easier," downloaded from https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html.
- Twitter (2018). "@Honda gives away a new car and \$100K to charity in instant win campaign on Twitter," downloaded from <https://marketing.twitter.com/na/en/success-stories/honda-gives-away-a-new-car-in-instant-win-campaign.html>.
- Wei, X., & Croft, B. (2006). LDA-based document models for ad-hoc retrieval. In *proceedings of the 29th annual international SIGIR conference*.
- Winkelmann, R. (2008). *Econometric analysis of count data*. Springer Science & Business Media.
- Winkelmann, R. (2013). *Econometric analysis of count data*. Springer Science & Business Media.
- Zadeh, A. H., & Sharda, R. (2014). Modeling brand post popularity dynamics in online social networks. *Decision Support Systems*, 30, 59–68.
- Zaman, T., Fox, E. B., & Bradlow, E. T. (2014). A Bayesian approach for predicting the popularity of tweets. *The Annals of Applied Statistics*, 8(3), 1583–1611.
- Zhang, Y., Moe, W. W., & Schweidel, D. A. (2016). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, 100–119.
- Zubcsek, P. P., Chowdhury, I., & Katona, Z. (2014). Information communities: The network structure of communication. *Social Networks*, 38, 50–62.
- Zubcsek, P. P., & Sarvary, M. (2011). Advertising to a social network. *Quantitative Marketing and Economics*, 9(1), 71–107.