DARDEN
BUSINESS PUBLISHING
UNIVERSITY of VIRGINIA

UVA-QA-0838
November 16, 2015

TWEETS, RETWEETS, AND ENTREPRENEURSHIP

The Tom Tom Founders Festival (Tom Tom), which began in 2012, was born out of a question: "How can live gatherings catalyze Charlottesville to generate new ventures, products, and visions of its future?"[1] Modeled after the popular South by Southwest (SXSW) festival held in Austin, Texas, each year, Tom Tom grew significantly in its first years. In 2015, the weeklong festival attracted over 25,000 attendees to the Charlottesville area, which had a significant positive impact on the local community. An economic impact analysis conducted in 2014 estimated a net local economic benefit of over $800,000 associated with the festival.[2]

Tom Tom's staff used Twitter in conjunction with other social media platforms to attract and engage their potential and actual audience. During the festival each year, the team noticed a significant spike in Twitter activity. In 2015, for example, mentions of @tomtomfest were up 466% during the 28-day period that included the festival. As a result, the team was interested in learning more about the characteristics of a tweet that made it more or less engaging, focusing on the use of keywords, such as "entrepreneurship."

The team wanted to make the festival's outreach and engagement with the entrepreneurial community on Twitter more effective. The number of times the festival's tweets were retweeted by others—presumably, a show of interest in the content—was an important metric to gauge the tweets' effectiveness. With a large dataset of past tweets on "entrepreneurship" collected from Twitter's application program interface (API), could the team identify the characteristics of tweets that tended to be retweeted more?

Understanding the Twitterverse

Twitter was founded in 2006 by Jack Dorsey, Evan Williams, Biz Stone, and Noah Glass as a social networking service that allowed its users to share short messages—up to 140

[1] http://tomtomfest.com/about/

[2] According to documents provided to case writer by Carolyn Zelikow, Assistant Director, Tom Tom Founders Festival.

characters—with one another. It quickly grew into one of the darlings of the tech world and expanded rapidly both in the United States and abroad. As more and more people began using the service regularly, Twitter became one of best ways to find out about what was going on in almost any part of the world.

While the site initially did not have any paid advertising, in 2010, Twitter began to support "promoted" tweets. For these tweets, brands or users could pay for their messages (tweets) to be displayed in the feeds of specific users. Such displays enabled brands to microtarget their ads to users based on their demonstrated interests.

Seven years after its founding, Twitter went public on the New York Stock Exchange on November 7, 2013, with the "TWTR" ticker. As of December 2014, Twitter had 284 million active users who were tweeting, on average, 500 million messages (tweets) per day. In the year since going public, revenue increased significantly, to $1.4 billion in 2014.

The platform was often used by professional athletes and performing artists such as Christiano Ronaldo and Taylor Swift to connect with their fans. Brands such as United Airlines, Target, and Coca-Cola also used it as an advertising and customer service platform to reach out to, and engage with, current and potential customers. In addition, professional and citizen journalists could immediately capture and relay developing news to a broad audience, as happened with the Boston Marathon bombings in 2013.

By April 2015, there were two primary ways in which users of Twitter could discover content on the service. First, when users logged into the mobile application or website, they saw their Twitter feed. A feed was composed of messages from users that they follow. These messages could either be tweeted or retweeted and were organized from most to least recent. A tweet was an original message composed by a user. Retweets consisted of another user reposting an original tweet, which could then be seen by the retweeter's followers. A user typically retweeted a message when he/she found the content engaging. Approximately only 1% of tweets are ever retweeted.[3]

Users could also discover content through search. A search could be conducted based on a user's Twitter handle (@) or a topic, denoted by a hashtag (#). A handle comprised the @ symbol, followed by an alias chosen by the person or organization. For example, noted entrepreneur and investor Elon Musk used the handle @elonmusk to reach his approximately 2.7 million followers and share information about his companies, included Tesla and SpaceX.

Hashtags were used to mark specific topics or events, such as #startup, if a person was interested in content related to startups, or #tomtomfest to follow tweets related to the music, arts, and innovation-focused Tom Tom Founders Festival that took place each April in Charlottesville. Users included hashtags in their tweets so other users can find tweets on a particular subject at a later date. Hashtags that were referenced frequently in a set period of time were said to be "trending" and would appear in a special "Trends" section on the left side of the homepage.

[3] http://www.quora.com/What-percentage-of-tweets-are-retweeted

Regardless of the specific aim (promotional, informational, etc.), the overarching goal of every Twitter user was for his/her content to be seen and engaged with repeatedly.

Accessing Twitter's API

While Twitter users saw a feed of tweets customized based on the specific people they chose to follow, behind the scenes, developers and data scientists worked with Twitter's application program interface (API). An API describes the set of routines, protocols, and tools used to build software programs.[4] For Twitter, the API not only consisted of the tweet, but also an assortment of metadata related to the tweet, which provided information on the user, source of the tweet, and location of the tweet (if available). Twitter provided access to its public API for app developers, which allowed them to track tweets based on a user or keyword, as well as publish tweets through the platform.

In order to begin working with the Twitter data, a developer would create a user profile within the Twitter developer system and obtain key codes to access the public API. Tweets could then be accessed via the Streaming API, which allowed developers to collect tweets over a continuous period by opening up a "stream", usually focused on a set of keywords.[5] There was also a REST API available, which pulled tweets as a snapshot in time, based on a specific keyword, time frame, and/or number of tweets requested. The information provided and its ease and speed of access varied between these two methods, so one had to choose the method most appropriate for the task at hand.

Even though Twitter provided free access to its public API, the volume of data provided through the service was limited. Twitter limited the number of queries and number of tweets a user could run at any given time. These were known as "rate limits" and varied throughout the day.[6] The REST API also limited the recency of tweets, usually only pulling tweets posted in the past 24 to 48 hours. Additionally, the request would not necessarily capture *every* tweet related to the specific keyword, so any data collection could be considered a random sample.

Exploring the Data

To inform Tom Tom's social media presence, the team decided to study the viral nature of similar content on Twitter. It collected tweets from a 24-hour period using the Streaming API and the keywords "entrepreneurship", "entrepreneur", "entrepreneurs", and "entrepreneurial". The data

---

[4] http://www.webopedia.com/TERM/A/API.html
[5] https://dev.twitter.com/streaming/overview
[6] More information at https://dev.twitter.com/rest/public/rate-limiting

from the Streaming API came in the JavaScript Object Notation (JSON) file format. In total, there were over 50,000 tweets in the file.

After some massaging and cleaning[7], the data were ready for further analysis. The dataset's variable names and descriptions are listed in **Exhibit 1**. At first glance, it appeared that there were several variables in the list that could potentially prove valuable in predicting the number of retweets.

One potentially valuable variable was "Status Source". It captured how each tweet was sent. Each tweet came with an embedded URL that identified if it was published through an iPhone, desktop, tablet, or a social media management system. Social media management systems were programs used to find the best time to tweet among other things. Their recommendations were based on algorithms that determined when other users would be more likely to see the tweet and reply or retweet. Some systems even allowed the user to program tweet rules that instructed the program to release tweets when certain keywords were trending.

In addition to the variables Twitter provided, there were others the team planned to engineer, such as "Handle Count". If other users were referenced in a tweet by their handle, "Handle Count" may indicate that the tweet was written by an experienced user or that the tweet was part of an on-going conversation. Either way, tweets with high "Handle Count" may be more likely to be retweeted. In order to isolate a handle, though, an analyst would need to understand how to extract a specific pattern within a string of text. RegExr was an open-source site that allowed a user to paste in snippets of text and to get back code for converting that text into a quantified variable.[8] This tool was used to engineer several of the new features, and could be used to create countless more.

Although not yet a variable in the dataset, the team planned to include each tweet's sentiment score. To do so, it decided to use the Hedonometer.org's happiness dictionary. With this dictionary, an analyst could determine the overall "happiness" contained in a tweet. Compiled by a team in the University of Vermont's Computational Story Lab, the Hedonometer.org's dictionary was one of the more robust and understandable dictionaries out there. It included more than 10,000

---

[7] The data were cleaned to prevent any data leakage. Data leakage results from information and/or variables in the raw dataset that could "give away" the dependent variable before forecasts of it are made and tested. In the case of Twitter, several of the provided variables and part of the tweet text contributed to this data leakage. For instance, because retweeted tweets began with "RT@[handle]", one needed to access the granular details of a retweet in order to pick up the text of the original tweet. Also, the only tweets included in the final dataset were those that were originally tweeted within the first 23 hours of the 24-hour collection period. The retweet counts included in the final dataset are the counts of retweets made within an hour of the time the orginial tweet was created.

[8] http://www.regexr.com/

scored words. The scored words were the most frequently-used words from Google Books, New York Times Articles, music lyrics, and Twitter messages.

Using Amazon's Mechanical Turk service, the dictionary's creators asked individuals on the service (who were called "turkers") to score each word on a nine-point scale from sad (1) to happy (9).[9] A word's final score was the average of the turkers' scores for that word. To measure a tweet's sentiment, one simply added up its words' happiness scores. For example, the tweet "Reading the wrap up email from @TomTomFest is inspiring. Congrats to the team on a great festival for CVille. #BetterNotBigger"[10] was run through the dictionary to determine its overall happiness score of 89.24. The tweet had several particularly happy words, including "festival" (7.7), "great" (7.88), and "inspiring" (7.34). On the other end, the lowest scoring word in the dictionary was "terrorist" (1.3).

The Challenge

The Tom Tom team had a content strategy in place for Twitter—focusing on entrepreneurship and innovation to encourage a positive conversation with its audience. Now it was ready to take the next step toward expanding the festival's reach and the positive impact it was having on the community. What was the best way to do that? What made a tweet go viral? Was there a better way to craft their tweets? Should it include particular words or topics, or mention particular people?

[9] http://hedonometer.org/about.html
[10] https://twitter.com/tomtomfest

Exhibit 1
TWEETS, RETWEETS, AND ENTREPRENEURSHIP
Dataset from Twitter API

Variable	**Definition**
ID_str	Unique ID; assigned by Twitter
Text	Text of the tweet
Created	Date and time the tweet was created
Status_Source	URL identifying how a tweet was published (i.e., iPhone, Hootsuite)
Screen_Name	Handle/Screen Name of the user who originally published the tweet
Latitude*	Latitude location of tweet
Longitude*	Longitude location of tweet
Followers_Count	Count of users following the user who published the tweet
RT_Count_in_TimeWindow**	Number of times the tweet was retweeted in a one-hour time period
Media_Manager***	Binary indicating whether a media manager was used to publish
Handle_Count***	Number of handles '@' that are in tweet's text
Created_Hr***	Hour of the day the tweet was created

* Location information is controlled by the users. The user must allow Twitter to access their location in order for this information to be embedded in the tweet's information. In most cases, users opt not to share this information.
** The original dataset was reduced just to those tweets that had a one-hour window of opportunity to be retweeted within our collection period.
*** These features were engineered from the original dataset.