THE WALL STREET JOURNAL.
WSJ.com

TECHNOLOGY | OCTOBER 1, 2011

Decoding Our Chatter

Want to monitor an earthquake, track political activity or predict the ups and downs of the stock market? Researchers have found a bonanza of real-time data in the torrential flow of Twitter feeds.

By ROBERT LEE HOTZ

When Virginia's magnitude 5.8 earthquake hit last August, the first Twitter reports sent from people at the epicenter began almost instantly at 1:51 p.m.—and reached New York about 40 seconds ahead of the quake's first shock waves, according to calculations by the social media company SocialFlow. The flood of messages peaked at 5,500 tweets a second.

Getty Images

Compared with information from cellphone records and social-media sites, Twitter texts are as timely as a pulse beat and, taken together, automatically compile the raw material of social history.

The first terse tweets also outpaced the U.S. Geological Survey's conventional seismometers, which normally can take from two to 20 minutes to generate an alert. The agency is now experimenting with Twitter as a faster and cheaper way to track earthquakes.

Never have scientists had so much readily accessible, real-time data about what people say. Twitter, the service that allows users to send text updates of up to 140 characters out to the public, publishes more than 200 million messages, or tweets, a day. Compared with information from cellphone records and social-media sites, Twitter texts are as timely as a pulse beat and, taken together, automatically compile the raw material of social history.

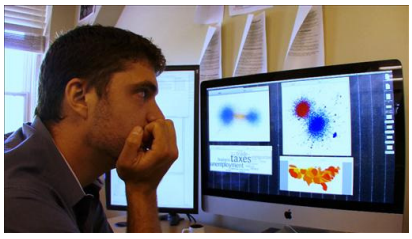
As Twitter's message traffic has grown explosively, so has the scientific appetite for the insights the data can yield. Dozens of new scholarly studies over the past 18 months by computer-network analysts and sociologists have plumbed the public torrents of data made available by Twitter through special links with the company's computer servers. This research has harnessed the service to monitor political activity and employee morale, track outbreaks of flu and food poisoning, map fluctuations in moods around the world, predict box-office receipts for new movies, and get a jump on changes in the stock market.

When the magnitude 8.8 Chilean earthquake hit last year, researchers found that on Twitter the truth often won out over misinformation. "When a rumor is true, it spreads faster," said computer analyst Barbara Poblete at the University of Chile in Santiago.

Ms. Poblete and her colleagues analyzed how survivors of the earthquake used the messaging service in lieu of more conventional communications that had been knocked out. They discovered that in the crisis, Twitter crowds reflexively sorted facts from falsehoods, exercising a collective wisdom on the fly. She found enough measurable differences in language, citations and posting patterns to devise a way to assess the credibility of Twitter texts automatically, with an accuracy of about 70%.

"The network itself can provide a filter for valid information," Ms. Poblete said.

All of this data is also proving to be valuable in the marketplace. Hundreds of social media, data-mining and financial-services companies now are paying a base rate of up to $360,000 a year for Twitter's information, according to executives at the two companies that are licensed to market it world-wide—Gnip Inc. in Boulder, Colo., and Datasift in Reading, U.K. "Twitter is protective of who has the data and where it is going," said Nick Halstead, chief operating officer at DataSift. "It is the ultimate customer research tool."



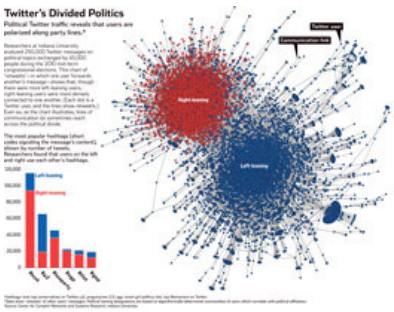
In an era of digital deception, scientists at Indiana University are using Twitter to investigate the nature of truth, lies and politics. WSJ's Robert Lee Hotz reports.

Though the practice is still experimental, Twitter data already have become a key variable in behavioral finance investment formulas. "The hedge funds are leading the way," said Chris Moody, chief operating officer at Gnip. Mr. Moody declined to name Gnip's financial customers. "They don't want anyone to know their secret sauce," he said.

The company does supply Twitter data to an investment firm in London called Derwent Capital Markets, which set up a $40 million hedge fund in May that openly uses a Twitter-based formula to guide its investment decisions.

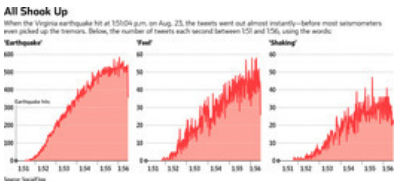
Researchers at Indiana University and the University of Manchester who developed the fund's technique say that they can reliably predict changes in the stock market by up to four days, based on the ups and downs of the national mood as expressed through key words in texts sent by 130,000 regular Twitter users.

Twitter's Divided Politics



Twitter's Global Moods

More photos and interactive graphics



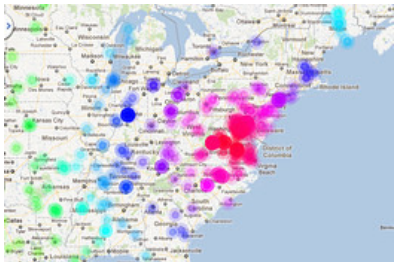
Tweeting the Tremors

'Omg earthquake!!!'

'Coworker: Was that an earthquake?? Me: Not sure, let me check Twitter.'

'Read about earthquake on Twitter before I felt it in Boston.'

'1:51 Earthquake. 1:54 First Earthquake jokes hit Internet. 2:05 Everyone is already sick of Earthquake meme.'



"We can make these predictions in real time, and I think it can be leveraged by a hedge fund to gain an advantage in the market," said Indiana computer scientist Johan Bollen, an adviser to the Derwent fund who helped to pioneer the sentiment analysis technique. "We have become more confident that this actually works."

After its first full month of trading in July, the investment firm announced that it had out-performed the Standard & Poor's 500 for that month, returning 1.85% while the index fell 2.2%.

Researchers led by Bernardo Huberman at Hewlett-Packard's Social Computing Laboratory have used Twitter to predict box office hits and flops. They successfully forecast the financial fate of 24 films, including "The Blind Side" and "New Moon," by analyzing the intensity of the word-of-mouth about them on Twitter. "We are interested in doing the same thing for products," said Dr. Huberman.

Other researchers remain skeptical of Twitter's purported predictive power.

This summer, for example, researchers at Wellesley College in Massachusetts examined the Twitter traffic during six close congressional elections last year, trying to see if the volume and emotional tone of the messages related to each race could have been used to predict the outcomes. In all, they analyzed a quarter million messages involving more than 60,000 people.

"Twitter did no better than chance," reported computer scientist Eni Mustafaraj, who led the research.

The military has recognized Twitter as a new battlefield for information warfare. In July, the Pentagon's Defense Advanced Research Projects Agency began exploring the possibility of a $42 million effort to detect online "persuasion campaigns" and "influence operations" aimed at spreading ideas through Twitter and other social media. The agency also wants to develop new technology for automatically "counter-messaging" adversaries.

"Changes to the nature of conflict resulting from the use of social media are likely to be as profound as those resulting from previous communications revolutions," said DARPA spokesman Eric Mazzacone in a written response to questions. "Adversaries may exploit social media and related technologies for disinformation."

At Southeastern Louisiana University, researchers reported that they could track influenza outbreaks by collating the rise in Twitter texts from people complaining about flu symptoms as effectively as more conventional public health reporting methods used by the U.S. Centers for Disease Control.

Unlike other instant-messaging systems, email, Facebook or Google, the personal information sent through Twitter accounts is public by default. Anyone with a free account can tap into the streams of conversation, merge themes or introduce new topics by employing short codes called hashtags, which are used to earmark subjects of discussion.

"With Twitter, you have a microphone, in effect, above all the millions of conversations that are going on during a day," said computer scientist Alan Mislove at Northeastern University in Boston, who uses the messaging service to track rumors, national moods and commercial brand information. "These pieces of information don't reveal much by themselves, but when you add them together they reveal quite a lot, and that's when it starts to get scary."

Last year, in an analysis of over 300 million tweets, Mr. Mislove and his colleagues found that people's moods follow consistent patterns over the hours of a day (with the highest levels of happiness in early morning and late evening) and the days of the week. The mood of each tweet was inferred by keywords like love, paradise and suicide. And, they found, people on the West Coast were significantly happier than people on the East Coast.

Researchers concede that their studies have some limitations. Twitter users tend to be younger adults, urban, more affluent and less likely to have children; they are not a cross-section of society as a whole. Still, researchers say, there is considerable diversity—demographic, national and cultural—among those who use the service, and it is possible to make meaningful generalizations from the flow of their messages.

No one is sure exactly how many of Twitter's 200 million or so registered user accounts are active at any one time and how many are dummy accounts. Twitter recently acknowledged that only half send messages. Some account holders aren't even human. Automated software programs called "bots," designed to spread advertising blurbs, run them.

A relatively small group of 20,000 users commands the most attention, researchers at Yahoo Research have discovered. They are neither the most prolific nor the most widely followed users, but the website links they recommend are more often repeated and shared by others. When it comes to focusing public attention, content matters more than celebrity, the studies suggest.

Scanning 580 million tweets over eight months, Stanford University researchers discovered that Twitter topics seemed to rise and fall in six distinctive patterns that could help to predict their popularity. At Cornell University, network analysts discovered that bad news appeared to fade fastest, weighed down by words with negative connotations. Good news more often floated to the top, buoyed in part by words with positive associations.

As Twitter markets its commercial data more aggressively, some scientists say their requests for access to Twitter's full data stream are being turned down more often. "Twitter has definitely become more wary about sharing their data," said computer scientist Jon Kleinberg at Cornell University.

Twitter executives declined to be interviewed about the company's sale and sharing of data. A spokeswoman said in a written statement that the company actively supports academic research—up to a point. Twitter is donating all of its message data to the U.S. Library of Congress, but it may be years before it is available and then only with restrictions on its use imposed by the company.

Many computational sociologists believe that Twitter offers a unique prism for studying communications across the political spectrum—and a rich source of strategic intelligence for targeting voters.

Researchers say they can easily predict a Twitter user's political leanings by looking at whose messages they relay to friends and followers and matching them to gender, location and other interests. The hashtag codes used to denote discussion topics give network researchers a reliable way to chart fluid political alliances. The researchers can also sort Twitter messages automatically by tell-tale keywords.

Twitter also has become a powerful political organizing tool. University of Michigan researchers pored through Twitter posts from 700 campaigns in the 2010 election and found that conservative candidates were more likely than liberal candidates to use Twitter to broadcast campaign messages. When it comes to Twitter, conservative activists were more organized, more in touch with each other, and more likely to stay on message.

The new messaging medium has also spawned a new form of political deception, in which campaign operatives marshal an array of dummy Twitter accounts to spread rumors or misinformation. Like form letters, robo-calls and push polls, these Twitter tactics are inexpensive, since user accounts are free, and can potentially reach many more people than traditional campaign attack ads.

By analyzing millions of tweets during recent U.S. elections and policy battles, researchers at Indiana University and other non-partisan computer analysts have identified dozens of cases in which activists orchestrated networks of dummy accounts, apparently operated by computerized scripts, to sway swing voters, influence pending legislation or promote a partisan cause by turning the popular messaging service into a political echo chamber of automatically re-tweeted texts.

"This is manipulation of social media, not to sell a product or steal a password, but to manipulate public opinion," said computer scientist Filippo Menczer at Indiana University's Center for Complex Networks and Systems Research, which monitors Twitter traffic to document such practices. "It is so cheap and easy. The incentives for abuse are huge."

They detected efforts to spam the system for political ends from both sides of the partisan divide. On the right, for example, they uncovered a pair of accounts that, mimicking the chatter of two politically active women, sent out more than 20,000 messages promoting Republican congressional candidates. On the left, they found 15 orchestrated Twitter accounts acting in unison to promote liberal immigration reform. A third account transmitted more than 15,000 texts fanning anti-Muslim sentiments, including links to a video of a beheading.

These prolific tweeters were most likely not real people, the scientists determined, but automated shams, based on the pattern and volume of the messaging. This sort of deception appears to be evolving faster than Twitter Inc.'s security measures can control them. The company forbids spam and efforts to mislead, confuse or deceive people.

In anticipation of the upcoming U.S. presidential contest, researchers at Indiana University have been working on ways to detect and defuse Twitter misinformation campaigns automatically. But the technology of Twitter is moving so quickly that detection efforts can barely keep pace. "People can game these systems and, in gaming them, they help bias the results of any data company," said social media analyst Danah Boyd at Microsoft Research. "It's a real challenge."

Pitting machine intelligence against human gullibility, researchers at the Web Ecology Project in San Francisco are using Twitter as a proving ground for advanced pre-programmed personalities called "socialbots" that can engage in extended conversations via Twitter by imitating the behavior of real people sending and receiving messages.

Designed to attract a large Twitter following, these code creations are constructed as an experiment in human-machine interactions, but the software could readily be turned to other purposes. "For good or for ill, you can get people to talk about a topic and potentially affect real-world behavior," said independent software developer Tim Hwang, who has been overseeing the effort. "If the bots are well-designed, they are undetectable."

In surreptitious tests online earlier this year, these socialbots fooled 300 unwary Twitter users. After refining their software, the group this month launched dozens of even more sophisticated Twitterbots, hoping to build relationships with thousands of unsuspecting users.

One Twitterbot from an earlier experiment—its account now disabled —masqueraded as a sports enthusiast. "I love going on adventures whenever I can find the time to dust off my passport," its biographical profile read. Its profile picture showed an exultant mountain climber.

"Once we launched it, it was fully on its own," said software engineer Greg Marra at Google in Mountain View, Calif., who helped to develop the bot as a college project. By design, "it would pick up a tweet from another user and parrot it. Completely unsupervised, it could produce a stream of plagiarized tweets."

During its nine months as an active Twitter user, it sent hundreds of messages about sports, sex, diabetes and the importance of online marketing. It attracted 1,538 followers, who apparently never realized they were in a relationship with a robot.

Network sociologists are worried that these newest contrivances may offer others a powerful way to manipulate people through Twitter on an even larger scale. "Doing this on Twitter with a thousand accounts or a million accounts is the next step," said Indiana University computer scientist Jacob Ratkiewicz.