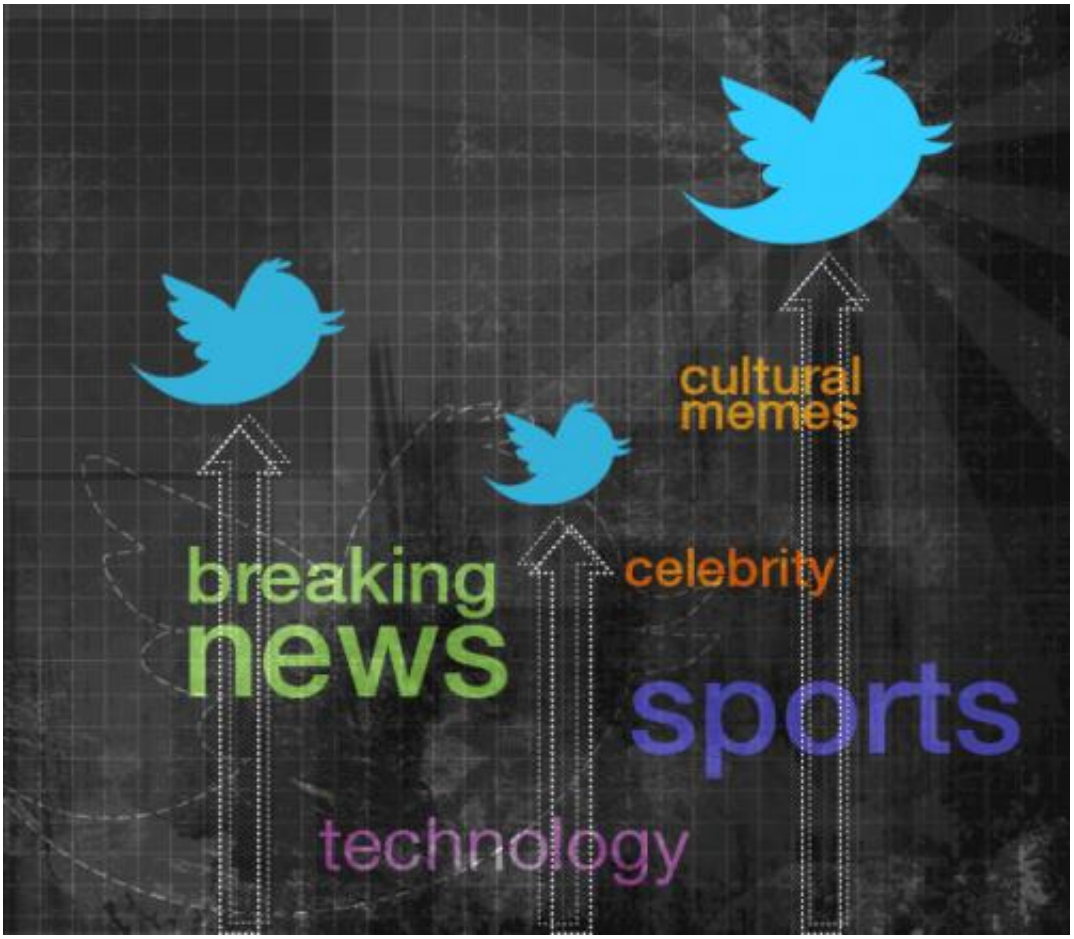
Predicting what topics will trend on Twitter

November 1 2012, by Larry Hardesty



Credit: Christine Daniloff

Twitter's home page features a regularly updated list of topics that are "trending," meaning that tweets about them have suddenly exploded in volume. A position on the list is highly coveted as a source of free publicity, but the selection of topics is automatic, based on a proprietary

algorithm that factors in both the number of tweets and recent increases in that number.

At the Interdisciplinary Workshop on Information and Decision in Social Networks at MIT in November, Associate Professor Devavrat Shah and his student, Stanislav Nikolov, will present a new algorithm that can, with 95 percent accuracy, predict which topics will trend an average of an hour and a half before Twitter's algorithm puts them on the list—and sometimes as much as four or five hours before.

The algorithm could be of great interest to Twitter, which could charge a premium for ads linked to popular topics, but it also represents a new approach to statistical analysis that could, in theory, apply to any quantity that varies over time: the duration of a bus ride, ticket sales for films, maybe even stock prices.

Like all machine-learning algorithms, Shah and Nikolov's needs to be "trained": it combs through data in a sample set—in this case, data about topics that previously did and did not trend—and tries to find meaningful patterns. What distinguishes it is that it's nonparametric, meaning that it makes no assumptions about the shape of patterns.

Let the data decide

In the standard approach to machine learning, Shah explains, researchers would posit a "model"—a general hypothesis about the shape of the pattern whose specifics need to be inferred. "You'd say, 'Series of trending things … remain small for some time and then there is a step,'" says Shah, the Jamieson Career Development Associate Professor in the Department of Electrical Engineering and Computer Science. "This is a very simplistic model. Now, based on the data, you try to train for when the jump happens, and how much of a jump happens.

"The problem with this is, I don't know that things that trend have a step function," Shah explains. "There are a thousand things that could happen." So instead, he says, he and Nikolov "just let the data decide."

In particular, their algorithm compares changes over time in the number of tweets about each new topic to the changes over time of every sample in the training set. Samples whose statistics resemble those of the new topic are given more weight in predicting whether the new topic will trend or not. In effect, Shah explains, each sample "votes" on whether the new topic will trend, but some samples' votes count more than others'. The weighted votes are then combined, giving a probabilistic estimate of the likelihood that the new topic will trend.

In Shah and Nikolov's experiments, the training set consisted of data on 200 Twitter topics that did trend and 200 that didn't. In real time, they set their algorithm loose on live tweets, predicting trending with 95 percent accuracy and a 4 percent false-positive rate.

Shah predicts, however, that the system's accuracy will improve as the size of the training set increases. "The training sets are very small," he says, "but we still get strong results."

Keeping pace

Of course, the larger the training set, the greater the computational cost of executing Shah and Nikolov's algorithm. Indeed, Shah says, curbing computational complexity is the reason that machine-learning algorithms typically employ parametric models in the first place. "Our computation scales proportionately with the data," Shah says.

But on the Web, he adds, computational resources scale with the data, too: As Facebook or Google add customers, they also add servers. So his and Nikolov's algorithm is designed so that its execution can be split up

among separate machines. "It is perfectly suited to the modern computational framework," Shah says.

In principle, Shah says, the new algorithm could be applied to any sequence of measurements performed at regular intervals. But the correlation between historical data and future events may not always be as clear cut as in the case of Twitter posts. Filtering out all the noise in the historical data might require such enormous training sets that the problem becomes computationally intractable even for a massively distributed program. But if the right subset of training data can be identified, Shah says, "It will work."

"People go to social-media sites to find out what's happening now," says Ashish Goel, an associate professor of management science at Stanford University and a member of Twitter's technical advisory board. "So in that sense, speeding up the process is something that is very useful." Of the MIT researchers' nonparametric approach, Goel says, "it's very creative to use the data itself to find out what trends look like. It's quite creative and quite timely and hopefully quite useful."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology