

## Getting in the mood

Jan Piotrowski

### Sentiment analysis has a big future

An urge to know what others will be up to next is part of the human condition. Soothsayers, fortune-tellers, stockbrokers—and publications like this one—have been catering to that obsession since mankind first began making plans for the future. Their record has been mixed. The biggest hurdle is the apparent unpredictability of individual behaviour. But if you knew the mood of all those involved, might a clearer picture emerge?

The problem is that those involved can number millions, and their thinking is tricky to tap. But a new breed of forecasters think it is becoming a little easier. They are using “sentiment analysis” to pick out the emotionally charged words and phrases which pepper online exchanges.

For instance, Johan Bollen, of Indiana University, Bloomington, has been trawling social-networking sites like Twitter for hints about people’s disposition—and trying to see how collective mood swings follow and, more important, foretell the course of events. Dr Bollen’s software has so impressed Derwent Capital Markets that the investment boutique has licensed it for one of its funds.

Kalev Leetaru, of the University of Illinois, has looked at the (still) more ubiquitous old media. He reckons he has found a way to forecast revolutions. He examined almost 4m articles

from the BBC’s *Summary of World Broadcasts* (SWB), set up by Britain’s authorities shortly before the second world war. Its aim was to scour publicly available foreign information—newspaper articles, television and radio broadcasts, and the like—for hints of attitudes towards the West, plus any other potentially helpful titbits.

Today, the SWB covers 32,000 sources in over 130 countries. Foreign-language reports are meticulously translated to capture as many vernacular nuances as possible. Crucially, SWB data going back to 1979 have now been digitised. This has allowed Dr Leetaru to recruit a supercomputer to

### It accurately fingered the area where Osama bin Laden was most probably hiding

compare the relative frequency of positive and negative words in millions of reports, arriving at a figure for each report’s emotional tone. These figures are then combined with place-name clues about which part of the world the reports concern to produce a global map of sentiment.

Words can, of course, be used ironically to mean their opposite. Irony has been a bugbear of sentiment analysis: computers, unlike most people, continue to be stumped by it. Yet the length of the SWB’s content makes that less of a problem than it is for, say, short tweets.

Moreover, by looking at seasonal variations in tone (caused by factors like availability of food), Dr Leetaru’s software can tell whether any changes conform to a cycle of collective mood swings which have not historically sparked unrest. Only if the souring of sentiment appears out of the ordinary would the model predict real trouble brewing.

The results are remarkable. Dr Leetaru’s map correctly indicated that resentment for autocratic rule was about to boil over in Egypt and Libya weeks before it actually did. (The SWB had too few articles about Tunisia for a reliable prediction.) It also accurately fingered northern Pakistan as the area where Osama bin Laden was most probably hiding.

So far Dr Leetaru has looked only at how his model would have fared with the benefit of hindsight. In 2012 there will be more forward-looking predictions. That, though, is just an unscientific hunch. ■

Jan Piotrowski: online science editor, *The Economist*



### 2012 IN BRIEF

NASA launches its NuSTAR satellite, the first that can make images of “hard X-rays” which may be able to map black holes in space

► vemurafenib, a new drug against malignant melanoma.

The problem is that identifying the relevant mutations is hard. Because evolution has come up with many anticancer mechanisms, for a cancer to take hold properly requires several mutations. That is unlikely to happen unless the mutation rate in a cell is abnormally high. A common factor in cancer is an overarching mutation in the DNA-repair mechanism. Lack of proper DNA repair allows mutation to run riot, increasing the chance that several anticancer genes will be affected, but also hiding those mutations in a plethora of others that have no relevance to cancer formation. The former are the needles that oncologists wish to identify. The latter are the haystack.

#### A tissue of lies?

Comparison of healthy and cancerous tissue from the same individual will show which genes have mutated. Comparison of the sets of mutated genes from different examples of the same tumour will show which mutations are coincidences, and which are causes. Not only

will this tell drug-developers where to concentrate their efforts, it will also test a theory that is gaining ground among oncologists. This is that classifying cancers by tissue type is fundamentally wrong, and that they should be classified by genetic mechanism, no matter where they occur in the body. Once the consortium has finished its work, it will be clear whether that is a good way of looking at things, or whether, by contrast, particular types of tissue usually become cancerous in their own, distinct ways, with only minimal similarities between one tissue and another.

As of October 2011, 39 laboratories had signed up to analyse 20 types of cancer. The rest should find labs soon. Though the last “i”s will not be dotted and “t”s crossed until 2015, by the end of 2012 the consortium’s data co-ordination centre in Toronto will be awash with results, and drug companies will be able to start wading in and picking out promising targets. Turning the new knowledge into treatments will still take a while. But the size of the task will suddenly be a lot clearer. ■