Clarifying the Social Media Blur

BY USING POWERFUL FILTERS TO EXTRACT KEY INFORMATION, LIBRARY PROFESSIONALS CAN MINE NOISY 'BIG DATA' AND HELP THEIR ORGANIZATIONS UNDERSTAND AND INFLUENCE STAKEHOLDERS.

BY DAVID MILWARD, PHD, AND GUY SINGH

For an increasingly large section of the population, social media are part of everyday life, both at home and at work. The numbers speak for themselves: Twitter has 100 million active users and 200 million accounts and transmits 230 million tweets per day (Sullivan 2011), while Facebook boasts 483 million active daily users and 845 million active monthly users (Facebook 2012).

The amount of data being produced may be daunting at first sight. How can such varied, noisy data be useful for specific applications? The answer lies in mining this information, which gives us the opportunity to capture "the world's population thinking aloud," discover consumer opinions, behavior and experiences, see how messages are transmitted and how people are influenced, and even tap into people's creativity.

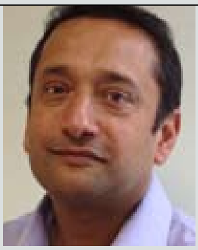
It turns out that the very size of the data is part of the solution. By imposing appropriate filters, we can find very useful data, but we need new techniques. Filtering the data using a traditional keyword search often works poorly on small documents (a tweet is limited to 140 characters). A keyword search can also return an enormous list of hits, so users either give up if the relevant result isn't among the top few returned, or they make do with whatever information they review first. Information professionals need to take a more rigorous approach.

Taking a Text Mining Approach

In the text mining world, the more documents there are, the better the results are likely to be. We can apply a wider range of filtering strategies than are possible in a keyword search, and we can also exploit terminologies to improve coverage. We can use the data itself to inform our search strategies,

DAVID MILWARD is chief technology officer and co-founder of Linguamatics. He has 20-plus years of experience in product development, consultancy and research in natural language processing, and is a pioneer of interactive text mining and its application to the life sciences. David has a doctoral degree from the University of Cambridge and was a researcher and lecturer at the University of Edinburgh.

GUY SINGH is senior product manager at Linguamatics. Over the last 20 years, he has held roles in most aspects of software development, ranging from research and development to product management and marketing. He has been instrumental in the development of innovative products for Internet, search and mobile systems for IBM, Oracle and Vodafone.



and duplicate results can be clustered together for much faster review.

Let's start with terminologies. A text mining approach can incorporate large-scale terminologies, providing hundreds of thousands of concepts and millions of terms. This allows users to look for tweets mentioning, say, *cancer*, and find terms like *tumor* and *carcinoma* as well as types of cancer, such as leukemia or Hodgkin's Disease. Terminologies also help unify information, since we can cluster results in terms of concepts rather than the actual words used (e.g., someone may refer to a Playstation 3 in one tweet and to a PS3 in another).

Text mining systems typically include natural language processing (NLP), which analyzes the structure of a sentence and breaks it down into distinct units to extract meaning. NLP can be used to find concepts that are not expressed as single terms. For example, the concept of "getting a flu jab" is expressed in thousands of different ways on Twitter, and these can be captured by a small number of patterns that exploit linguistic structure.

NLP can also be used to find associations between people, products, genes, and diseases as well as sentiments concerning them. In particular, it can find precise relationships and the direction of any given relationship (e.g., one product preferred over another, or a protein phosphorylating another protein). Relationships can also be chained together to suggest potential mechanisms of action—a chemical-to-gene association mentioned in comments about a conference presentation can be linked with a gene-to-disease relationship from the scientific literature. This use of text mining to generate new hypotheses and new knowledge from multiple documents is sometimes termed *text data mining* (Hearst 1999).

Agile text mining, also known as interactive information extraction (Milward et al. 2005), mixes traditional text mining with search technology to allow us to interactively develop queries, just as we might develop a keyword search strategy. With agile text mining, however, we also have the ability to create arbitrary queries mixing keywords, terminologies, regular expressions, co-occurrences, and NLP.

Using agile text mining techniques, a search can be refined systematically, since we can discover as well as search. We can mine the data to find the kinds of terms and abbreviations people actually use. If particular terms are too noisy, we can see how to refine the context—for example, by finding the most frequent words before or after the terms to find good candidates to include or exclude.

Finally, text mining allows us to cluster information according to the message being expressed or new information being provided. This clustering allows us to quickly skip less relevant results and spot duplicates. It is a much faster and more efficient way to analyze results than reading each document in which "hits" occur.

> With social media, you are tapping into unadulterated thoughts or a discussion thread. It is not an intrusive approach—no one is put on the spot for an opinion.

Applying Text Mining to Social Media

Our firm, Linguamatics, has undertaken a number of text mining projects in the past few years using Twitter as a data source. These projects have involved the following tasks:

- Researching consumer opinions on products, companies and people;
- Tracking consumer behavior and identifying how particular sections of the population are being influenced;
- Conducting competitive intelligence on what people are saying about competitors and their products;
- Capturing creative suggestions people are making; and
- Identifying how and where messages are being distributed and the key opinion leaders in different communities.

An NLP-based text mining system can find out what people are saying about a subject, not just that a certain subject was mentioned. It can accurately determine whether the opinion is positive, negative or undecided, and it can help with ambiguous product names such as *orange*.

Let's consider some generic examples (but note that, typically, much of the terminology needed for sentiment is specific to the product or issue being discussed—for instance, *soft* may be positive for a car's plastics and possibly its ride, but not its brakes).

Even in generic cases, there is some subtlety. Let's start with some simple positive cases:

- I really like Product X.
- Product X is great.
- I prefer Product X.

We may want to distinguish actual users of products from the general population, who are often discussing commercials for a product rather than the product itself. To identify users, we can look for constructions such as the following:

- Second day of using Product X.
- I stopped using Product X.
- Product X helped me lose weight.

We also need to distinguish actual opinions from conditional ones:

- Product X should be effective.
- If Product X works, I will buy it.
- I hope Product X is good.
- Do you think Product X will work?

Similarly, we need to ensure that positive words such as *like* are being used in the correct sense and that a word like *prefer* is referring to Product X

Agile text mining provides us with a powerful methodology to clarify what might otherwise be a blur of social media content.

rather than a competing product:

- Product X is like Product Y.
- I prefer Product Y over Product X.

Finally, we need to distinguish negative contexts that might reverse the polarity of words such as *like* or *work*:

- I don't like Product X.
- Product X didn't work.

We can use a variety of techniques to segment different users, or we can group them together. This approach can be used to determine whether different groups (1) have different behaviors and opinions or (2) are influenced in different ways. The preceding example of distinguishing users of products is just one scenario; other ways to segment groups are defined in the following paragraphs.

Similar opinions. Like-minded people can be clustered if they express similar sentiments about a particular subject.

Geographic location. Segmentation can be based on where users are from or where their mobile device was located when they tweeted. Specific locations can be mapped to wider areas, such as countries or regions of the world.

Time frame. Segmenting comments according to the time they were expressed is often important. Time slices may range from minutes (e.g., while users are commenting on a live event) to days or months.

Finally, it's useful to identify the movers and shakers (key opinion leaders) in social media in particular topic areas. These can be particular people, but also organizations or Websites. In addition to looking at followers in Twitter, we can also discern the effectiveness of messages based on the extent to which the messages are re-tweeted to others.

Why Not Use a Focus Group?

How do these techniques compare with assembling a focus group or conducting a telephone poll? Social media reflect people's opinions without asking questions explicitly. With social media, you are tapping into unadulterated thoughts or a discussion thread. It is not an intrusive approach—no one is put on the spot for an opinion. It is also scalable and quicker, because you don't need to assemble a group of people. They are already there, all the time.

Focus groups and polls have the advantage of being able to provide a carefully balanced selection of the population, not just those with the loudest voices. However, this can also be a disadvantage because it can miss the dynamic and viral nature of opinion. Key opinion leaders do have an effect: a single joke or comment can change perceptions of an issue. We often need to monitor these fast-moving flows of opinion and must be prepared to intervene if a false rumor needs to be countered.

To illustrate how text mining can help provide clarity and noise reduction in Twitter, we developed a number of case studies with Royal Holloway, University of London. These studies used Linguamatics' Interactive Information Extraction text mining platform, I2E.

Case Study: Sony PS3 Online Failure

In March 2010, a major fault was reported that stopped Playstation 3 users from accessing the online network. We used this event to track user sentiment and also look at whether there were any emerging common theories on the cause of the problem. By monitoring the Twitter traffic, we picked up on the major issue that Sony was not providing updates to the many bewildered users who could not figure out what was happening.

We were able to cluster all of the common opinions, even though different language was being used.

- RT @DanAdams85: Come on Sony! 10 hours and no update! Just some info on what is going on would be nice.
- #PS3 Still no answer from Sony, latest update dates from 12 hours ago. Seems like meetings are happening in Sony's HQ.
- @Sony But @SonyPlayStation hasn't said anything for 14 hours. This isn't looking good for the brand ... And I'd like to continue playing ...

We were also able to look at trends arising from different theories about the cause of the problem. Initially, "hacking" was commonly suspected, but this was eventually replaced by the true cause—a millennium clock bug-style error.

Case Study: Avoiding the Flu Pandemic

During 2009, a flu pandemic (sometimes referred to as swine flu) was reported as breaking out. We conducted a study to identify who was planning to be vaccinated and what or who was influencing them.

Using NLP enabled us to distinguish between those who just mentioned vaccines and vaccinations from those who expressed a preference about being vaccinated. This use of NLP allowed us to filter out most of the noise around this issue (including spam) and segment the population into those getting the vaccine and those not getting the vaccine. Network visualization allowed us to draw some interesting conclusions about how different groups were being influenced.

Figure 1 represents a subsection of the whole picture. In the green diagram, we see social media users who are getting the vaccine; in the

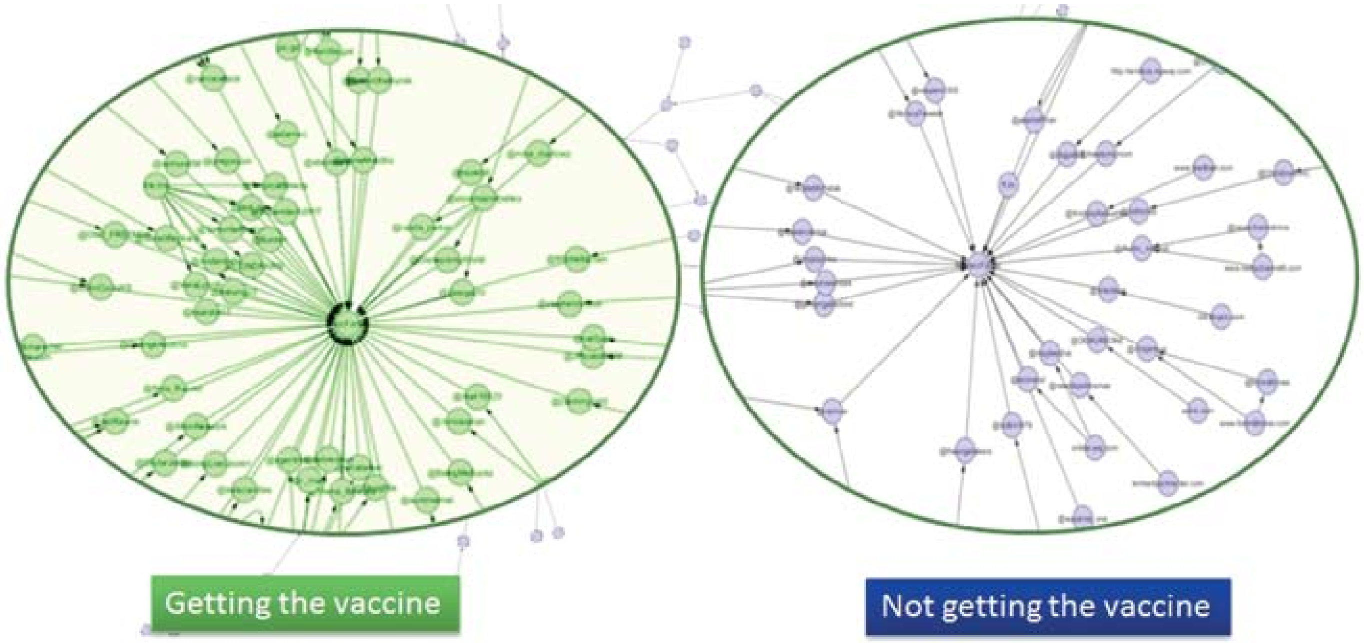


Figure 1: Case study of social media data relating to the 2009 flu pandemic.

center of the diagram we see a particular tweet (a reference to flu.gov) that is influencing many other users. Likewise, in the blue diagram, we see that a particular tweet—about a natural remedy Website—seems to have a number of others clustered around it. Finally, between the green and blue are a number of interconnecting nodes representing users who have a foot in both camps.

Case Study: 2010 U.K. Elections

During the run-up to the 2010 U.K. elections, Linguamatics conducted a project to not only monitor mass opinion and the sentiments of the electorate, but also to see if it was possible to use Twitter opinion data to predict the eventual result. The analysis centered on the three televised debates, each lasting 90 minutes:

- 15 April 2010 (ITV)
- 22 April 2010 (Sky)
- 29 April 2010 (BBC)

Approximately 567,000 tweets from 130,000 Twitter users were analyzed during this period. For each debate, positive sentiment toward each leader was measured using natural language processing.

The results produced by Linguamatics corresponded quite closely with opinion polls published by national TV networks and newspapers, which provided some verification of the integrity and accuracy of the data. The results were re-published on a number of Websites, including the BBC's.

By combining the results from all three debates, it was possible to detect trends in the candidates' popularity and predict who would become the next prime minister. The final result of the general election corresponded closely to the trend analysis created by Linguamatics.

A Powerful Methodology

Due to its popularity and pervasive use, social media cannot be ignored. But given the amount of social data, its continuing growth, and the amount of noise relative to useful information, we cannot rely on traditional methods such as keyword searches to extract what we need. The presence of informal speech, colloquialisms, slang, and sarcasms only adds to the challenge, making it even harder to find the right information.

In this article, we have discussed how agile text mining is well suited to dealing with this kind of data thanks to its ability to find information regardless of how it is expressed, reduce noise by looking at linguistic context, and cluster and synthesize information. This provides us with a powerful methodology to clarify what might otherwise be a blur of social media content, thereby providing efficient access to an increasingly important knowledge source. **SLA**

REFERENCES

Facebook. 2012. Company Information Factsheet.

Hearst, M.A. 1999. Untangling Text Data Mining. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 20-26 June.

Milward, D., M. Bjäreland, W. Hayes, M. Maxwell, L. Öberg, N. Tilford, J. Thomas, R. Hale, S. Knight, and J. Barnes. 2005. Ontology-based Interactive Information Extraction from Scientific Abstracts. *Comparative and Functional Genomics*, 6(1-2): 67-71.

Sullivan, D. 2011. Twitter CEO Dick Costolo's 'State of the Union' Address. *Search Engine Land* (blog), 9 September.