



Social Media Monitoring & Insights

Alkis Papadopoulos, Vice President Advanced Technologies

May 2011

© Voxco 2011

Can Social Media Monitoring achieve the ambitious results promoted by vendors today?

The answer depends greatly on the degree to which existing solutions or systems are weighted towards analysis and insights versus monitoring.

A Social Mining Primer

As evidenced by reports published regularly by the Altimeter Group, Forrester's Listening Platform, research reports, etc., and the rapidly growing number of Social Media Monitoring companies, it is safe to assume that we have passed the stage of asking whether or not monitoring conversations on the web are a nice-to-have or an outright necessity. Basic usage figures (2010) illustrate eloquently how prevalent the web and social media in particular have become:

...we have passed the stage of asking whether or not monitoring conversations on the web are a nice-to-have or an outright necessity.

- Wikipedia – more than 3.5 million subjects in English alone
- YouTube – more than 100 million videos
- Blogs – more than 150 million
- Twitter – more than 175 million users
- Websites – more than 250 million
- Facebook – more than 500 million users
- Internet users – approximately 2 billion

An organization wishing to engage more directly with its customers, communities, members, constituents or political groups can succeed by extracting key concepts and thoughts contained in their dialogs and contributions in multiple web media. This can be done through the following process:

First gather data that already exists on the web, sift the poor or irrelevant content from that which actually carries useful information. Then successfully extract key concepts and thoughts from said content and determine as efficiently as possible what call to action, if any, is required. Finally, route the call to action to the appropriate department, division or resource in a timely fashion.

In this regard, market expectations have matured as more social media monitoring products have appeared.

Beyond the above requirements, is the ability to successfully dissect the glut of feedback, in order to quickly gain real insight into the nature of problems or feedback reported [...].

The typical features expected from such products are:

- Central, comprehensive management of brands, competitors and topic tags.
- Restriction by languages and countries.
- Comprehensive sources (many tools do not scan all sources and eliminate tweets).
- Dashboards with Key Success Indicators (KSIs) as well as filters and comparison options (mentions, reach, share of voice, etc.).
- Sentiment analysis, demographic information where possible, identification of influencers and important topics.
- Filterable lists of contributions with information on range, user profiles etc.
- Historic data for retrospective analyses.
- Workflows and pragmatic interfaces for existing CRM systems.
- Direct engagement options.
- E-mail alerts (particularly in the context of crisis management).

Beyond the above requirements, is the ability to successfully dissect the glut of feedback, in order to quickly gain real insight into the nature of problems or feedback reported, and automate relevant calls to action specific for each problem discovered. Successfully mining this vast set of data is key, but approaches to extracting information from this data vary significantly. Particular attention is required when proceeding with what ultimately remains qualitative analysis from open sources that are subject to spam, offensive material and outright gaming.

Vetting of sources

Identification of key domains or websites where a company or organization believes there is useful interaction between people will help to ensure that content that is being analyzed is truly useful. This can include customer review sites, online communities or forums that are hosted by the company itself, mainstream news and blog sites, technical forums (where applicable) and of course prevalent social networks and micro-blogs, though these must be subjected to spam and offensive content filtering prior to analysis.

Content relevance

It is all about the search results. No matter how one slices the problem, in order to retrieve the data you are going to be analyzing, you will be writing a query or series of queries in order to get at the relevant data. Ultimately, your analysis can only be as good as the data you are analyzing. If the underlying Search Engine and relevance assessment models are not perfect, your analysis will suffer significantly, and reliable insights and truly actionable information will be difficult to obtain at best.

Sentiment analysis reliability

Truly accurate sentiment analysis continues to be elusive for several very good reasons. As a Natural Language Processing exercise, sentiment analysis must stand upon the foundation of a number of prerequisite steps that are not simple to achieve without certain error rates. And as we move through each step, error rates are being compounded. Key steps are:

...sentiment analysis must stand upon the foundation of a number of prerequisite steps that are not simple to achieve without certain error rates.

- Identify the language of the text to be analyzed.
- Remove boilerplate and unwanted noise, and extract the structure of actual user posts, articles or blogs.
- Identify (when applicable) within this structure, paragraphs and sentences.
- Separate sentences into their individual concepts.
- Identify the grammatical categories and sentence components such as subject group, noun-phrases, adjectives, verbs, adverbs, etc.
- Identify stylistic features such as punctuation.
- Extract and tag key concepts (semantic analysis as opposed to simpler key word extraction).
- Identify sentiment or emotion markers.
- Categorize the text being analyzed into categories that correspond to the most important and prevalent semantic tags.
- Classify sentences according to sentiment.

Voxco's Social Media Monitoring & Insights

Contrary to the bulk of vendors in the Social Media Monitoring space, Voxco has taken a natural language processing approach. This combines semantic analysis based on industry specific dictionaries with machine learning algorithms that are attempting to identify meaning and present results in context. We believe this is the only way to enable our customers to reliably mine data for key concepts and thus serve a wide range of opinion mining needs. But much more importantly, to assist in reducing the time that must be spent analyzing data in order to draw reliable conclusions, and obtain actionable information. Many opinion mining and social media monitoring solutions available today do a good job of collecting data, and monitoring for occurrences of key words (brand names, product names, etc.); but they leave the end-user with the task of manually figuring out how to proceed with analysis, and, from this flood of data, attempt to understand what the appropriate calls to action are within their organization.

Voxco's Social Media Monitoring and Insights has been designed and developed to not only ensure that users can collect the data, but also to assist in shaving costly hours off of the more complex data analysis phase.

Productivity and process

Voxco's Social Media Monitoring and Insights has been designed and developed to not only ensure that users can collect the data, but also to assist in shaving costly hours off of the more complex data analysis phase. This solution can thus truly help users gain actionable insight from analysis of feedback about their products and/or services. It also enables them to monitor the impact of recent public relations and marketing campaigns, and pursue targeted intervention tactics in order to redress perceptions, if and when they or their products are subject to criticism online. In order to do so, Voxco's Social Media Monitoring and Insights completes a series of steps that make it possible to get to the most important concepts found in individual users' feedback. Equating individual user feedback with a document, it is necessary to:

- Accurately identify the language of feedback if data is from open sources.
- Remove boilerplate for web content and effectively extract only the actual text written up by a user.
- Identify document structure, so as to separate out individual paragraphs and, more importantly, determine sentence boundaries within paragraphs.

The end result is a clean set of sentences that are tagged as belonging to a given instance of user feedback, with relevant concepts, entities and sentiment extracted for further analysis [...].

- Tokenize sentences (separation of sentences into its individual words).
- Proceed with partial syntactic analysis and part-of-speech identification (identification of grammatical categories such as subject group, noun-phrases, adjectives, verbs, adverbs, etc.).
- Identify stylistic features (titles, text formatting, accentuation through punctuation, etc.).
- Determine the semantic descriptor of each document.
- Determine the sentiment associated with each sentence.
- Identify important links between semantic metadata and sentiment metadata.

The end result of these steps is a clean set of sentences that are tagged as belonging to a given instance of user feedback, with relevant concepts, entities and sentiment extracted for further analysis by our Social Media Monitoring and Insights product. Where available, additional metadata associated with user feedback is stored along with the above, specifically:

- Language.
- Date and time stamp.
- Source of the user feedback (a URL for example).
- User identification (for example, a user's nickname on a blog or forum).
- User demographics (when from Enterprise Content sources only and not from the web where such information is very sketchy at best).

The identification of Key Concepts, Named Entities and Emotion Markers uses a combination of proprietary unsupervised and semi-supervised machine-learning techniques in order to optimize classification of sentences into the categories that are as precise as possible. The subsequent assignment of a sentiment flag to each sentence, also relies on proprietary machine learning algorithms. Successive pruning of concepts based on relevance feedback from the customers themselves, as well as filtering for topics specific to the industry or sub-industry of interest based on industry-specific dictionaries, help to ensure that the only the most salient topics emerge accompanied by an assessment of how emotional about these topics are the people who have written about them.

The value is in insights, not mining

Why go to these lengths, when we could more simply limit our efforts to identifying keyword patterns in social media feeds? The steps outlined above result in Voxco's relevance assessment model being the best applicable to text based information, and ensure that clients can get to the most interesting nuggets of information from user feedback.

[...]the system helps clients pre-analyze user feedback for Key Concepts, Named Entities and Sentiment, to ensure that they are retrieving the most useful and relevant information [...].

When setting up feedback studies, making use of our very effective relevance assessment model, the system helps clients pre-analyze user feedback for Key Concepts, Named Entities and Sentiment, to ensure that they are retrieving the most useful and relevant information about the topic or topics they have chosen to study. They can then obtain as complete and relevant a collection of user feedback as possible, for the topic at hand, and subsequent insights will be that much more interesting and useful.

Categories, Concepts and Entities

In retrieving information, the system proceeds with a number of pre-computations attempting to establish links between Key Concepts, People names, Organization names, and corresponding sentiment. The most relevant and prevalent of the above are labeled as categories and these help to guide customers through the analysis steps. Then we cross-reference categories to other categories to establish the context. This facilitates individual users to gain insights into:

- Who are the main contributors?
- Where their expertise lies?
- What is the sentiment, to determine whether the feedback is a complaint or praises the topic at hand?

Figure 1 shows how the most prevalent categories are presented with the terms or concepts that describe why they are positive or negative and what specifically is being praised or criticized. Figure 2 illustrates the breakdown of users per category, giving clear indication of the main ideas and concepts evoked by each user, the frequency of their posts or write-ups and recent distribution of sentiment.

With the approach developed at Voxco, we are striving to help customers reduce the most time-consuming part of social media monitoring projects, namely the analysis step.

Conclusion: do not settle for monitoring

This ability to show the context of the topics evoked is critical. It depends on the detailed natural language processing that Voxco's Social Media Monitoring & Insights brings to the table. This makes it feasible for customers to draw actionable conclusions then and there rather than spend time trying to sift through data that is difficult to read and interpret given that it is often limited to context-less keyword tracking. With the approach developed at Voxco, we are striving to help customers save the most time consuming part of social media monitoring projects, namely the analysis step. Our proprietary algorithms rooted in robust semantic relevance assessment, and deriving the context of all categories (Key Concepts, People, Organizations, etc.), are designed to help obtain answers to questions, gain insight where insight is elusive and ultimately save you time and money. Why settle for monitoring only when, with Voxco you can also get to actionable insights from all user feedback?