Moving from Measurement to Meaning in Text Analytics:
Tools to Wring New Insight on Buzz & Brand

By Michael Kelly, Ph.D.

Opportunities and Challenges as Vast as the Data

The surging availability of internet comments about brands, products, events, and people has spurred extensive use of automated text analysis to gauge market sentiment. Search algorithms can comb the Web for online product reviews to detect overall impressions, and in the service of ongoing market measurement, companies can develop sentiment scores by routinely monitoring Twitter, Facebook, or other social media channels to track brand health or gauge the impact of marketing activities.

With the computational power to analyze extremely large bodies of text, there is also mounting conviction that the results can truly be useful. Still, a number of significant measurement challenges remain, inspiring cautious interpretation and creative spade-work to improve our techniques. To illustrate the limits of sentiment analysis, the word "long" is positive in the phrase "the phone has a long battery life" but negative in "the video took a long time to stream." To obtain valid readings, analytic algorithms may need to infer context – a far higher hurdle than simply tallying expressions of sentiment.

Until such measurement issues are resolved, the potential for text-based predictive modeling to inform business decisions seems tantalizingly close but still just out of reach, leaving us with provocative questions about the ultimate meaning of our measures:

- *Can product sentiment scores automatically calculated from online reviews ever deliver a trustworthy forecast of sales?*
- *Might text patterns, like "dialects," ultimately distinguish market segments that vary in their propensity to purchase a particular product and respond to particular messages?*
- *Can brand strength be assessed not only in text that is explicitly evaluative, but even in common, apparently neutral remarks tapping new reservoirs of insight from web commentary?*

Size Matters

Encouragement about prospects for moving from measurement to meaning comes from case studies that reveal close linkages between text and attitudes/behaviors in simple analysis powered by sheer volumes of data. For example, levels of positive sentiment in Twitter feeds referencing President Obama or the economy have been shown to track closely with polling data on those

same topics, even when sentiment-capture and coding are based solely on the presence of positive or negative words ─ regardless of syntax or adjacency. In an even more "stripped down" analysis, researchers from HP's Social Computing Lab were able to predict box office revenues for a movie's opening weekend extremely well based on just *frequency* of Twitter mentions during the week before opening night without analysis of sentiment. Mere "buzz" proved to be a potent predictor.

The effectiveness of such basic analyses reflects the power of extremely large sample sizes - even samples that may not be perfectly representative of the broader US population. Despite considerable tweet-to-tweet variation, 3 million tweets from 1.2 million distinct Twitter users average out idiosyncrasies and allow detection of a predictive aggregate signal.

'Waste Not, Want Not': Picking Up Meaning from the Cutting Room Floor

Cases studies such as these highlight the potential power of text analytics for market research and, in fact, have reassured some skeptics that Big Data can reliably portray the big picture. But, as in many areas, we can do even better through data conservation. Here's why and how.

Some standard practices in the field (for instance, stripping out common "function words" like "the" and "your") will certainly improve computational efficiency, but they can also deprive us of useful insights ─ particularly when the goal is to move beyond sentiment metrics to more sophisticated inferential modeling. *As a case in point, models can predict the gender of a blogger with 72% accuracy using just function word distribution.* That's because men are more likely to employ determiners like "the" and "a" whereas women are more likely to use pronouns like "he" and "she" – linguistic patterns that reflect the propensity of men to discuss concrete objects and women to talk about social relationships. Rather surprisingly, "content" words like "football" boost modeling accuracy by only around 4%. This is an important finding since models based entirely on the very high frequency of function words can be developed and evaluated more efficiently than models that use content words, any one of which occurs much more rarely.

Here are some practical applications. Businesses can use them to develop models to inform market analyses and guide corresponding actions. One might, for instance, test whether women's sentiment scores better predict market success than men's, and thus discern who is ultimately "in charge" in a given purchase category, or use other demographic segment assignments to develop more sophisticated predictive models.

'This and That' in Text Analytics: Word Sequence Also Matters

At **NA**XION, we have been exploring the potential utility of a specific type of function word for text analytics ─ namely *conjuncts* like "and" and "or" that link together or "conjoin" two words into a single phrase. Conjuncts are potentially valuable terms because they are very common. *Even more important, though, it turns out that they provide important insights into speaker attitudes, based on order of mention.* Consider, for example, the following conjunctive phrases: "good and bad," "pleasure and pain," "love and hate," and "plus and minus." Each of these phrases contains a positive word and a negative word. Although both word-orders are equally acceptable from a grammatical standpoint, people tend to mention the positive word first. An analysis spanning two

centuries of digitized English text found that writers use “good and bad” about ten times more often than “bad and good.” Conjunctive phrases exhibit other similar patterns in which words connoting strength or power are routinely positioned first rather than second (e.g., “strong and weak,” “active and passive,” “leader and follower”). Thus, word-order patterns implicitly signal a person's attitudes toward what's mentioned in the conjunct giving clues about perceived value and strength.

Brand Name Sequence and Brand Strength

Let's extend our discussion of word-order patterns in conjunctive phrases to the marketplace. Brands often appear in conjuncts like “Apple and Google” or “Google and Apple,” such as in these, without any explicit semantic cues as to sentiment or priority:

> *“Both Apple and Google have their pick of students from any universities in the country.”*
>
> *“Google and Apple are rivals.”*

In word pairs like “Apple and Google” versus “Google and Apple,” the two brands are inevitably *competing* in the mind of the speaker for the halo implied by that first position. *Significant skews in word-order thus provide an implicit indication of brand strength in the marketplace, even when the speaker omits sentiment words and has no explicit intent to convey valence.* As case in point, our analysis of blogs over the past year finds that “Apple and Google” is more than twice as common as “Google and Apple,” suggesting that Apple holds primacy in share of heart and mind in the digital space.

Attention to word sequence is nothing new; in traditional studies of awareness and brand “first mention” is a common statistic. Brand order in natural text starts with a similar metric but offers the potential for richer, contextually embedded analysis and modeling given the huge volume of text now available. A much wider range of hypotheses can therefore be generated and tested efficiently, offering the potential to link words and actions in the market much more closely.

Putting Text into Larger Context

There is, as always, a note of caution to be raised. As we proceed with analyses of conjuncts or any other specific structure in the varied domain of text, interpretation of findings will need to be grounded in the broader market context, lest we draw incorrect inferences. In the smartphone space, the sequence “iPhone and Android” dominates “Android and iPhone” by almost 9 to 1 yet, on the most *bankable* measure of brand strength – market share – Android leads. This is a reminder that many market characteristics (e.g., infrastructure, distribution, and characteristic business models) can disrupt, or at least complicate, the link between sentiment and purchase behavior. And it underscores the fact that text analytic enhancements will need to be integrated with multiple streams of data (and their corresponding models) to discern if, how, and where words foretell actions. Integrative approaches like these, which have proven their worth in various market models developed over the years, will help to ensure that text is not taken out of “context.”