From Unstructured Text to Valuable Insights: Leveraging Text Analytics to Meet Competitive Intelligence Needs

By Tom H. C. Anderson, Anderson Analytics

Currently over 80 percent of all information is stored as text. Unlike structured databases, which allow competitive intelligence (CI) practitioners to easily identify patterns or trends, the unstructured format of text can prove difficult and time-consuming to analyze. Recently, several software companies' applications have transformed text into a more usable structured form, allowing many previously unavailable information sources to be included in competitive intelligence or market research programs.

Several Fortune 500 companies have realized that companies who first leverage this new technology stand to gain a considerable information advantage over their competition. Therefore, their independent efforts have not been publicized, and CI practitioners have surprisingly little information available to them on how and when to best make use of this technology.

This article defines text analytics, describes the techniques used behind the text analytics process, and illustrates the process itself. The article also reviews potential obstacles faced by individuals who use text analytics and shares some of the best practices developed by Anderson Analytics.

BACKGROUND ON TEXT ANALYTICS

Text, whether obtained through writing or transcribed from speech, is the primary method to record our thoughts, emotions, opinions, and beliefs. Information professionals collect attitudes and reactions through language, whether through

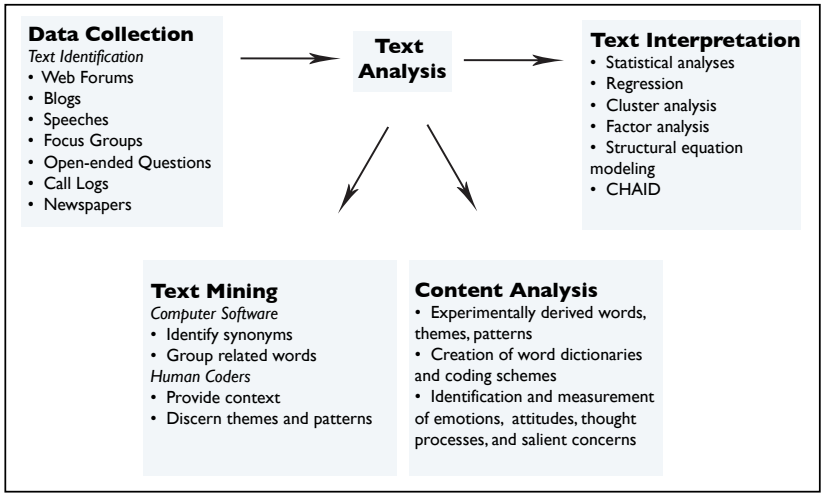

Figure 1: Text Analytics Process

primary research (such as self-reported surveys, focus group discussions, or one-on-one interviews) or through secondary research (such as articles, financial statements, and business news).

Text information can come from online sources ranging from internet web pages, discussion forums, blogs, and emails as well as from traditional media sources, such as newspapers, annual reports, letters, advertisements, speeches, interviews, and focus groups. By applying audio transcription, call center logs and a host of other audio materials can be sources for text analytics. The sheer variety of text sources is simply daunting.

The internet holds a nearly infinite amount of opinions and thoughts expressed primarily through text. Like most aspects of the web, this information is vast, convoluted, and unstructured. Understanding what consumers, competitors, or opinion leaders actually think about a specific topic, product, or brand is extremely difficult—any type of communication has an abundance of "noise" or irrelevant information, either in speech or written form.

The difficulty of gathering relevant opinions and thoughts increases when you want to garner useful information about people's attitudes toward a product or concept. The internet is a potentially powerful source for obtaining text information about a market group (either established or potential) that is unsolicited—potential respondents have already presented their attitudes and thoughts without the prompting of a survey or recruitment process. But how can you develop an information-gathering strategy based on unstructured or obscured information?

Text analytics provides quick and massive content identification and pattern recognition that can help you predict consumer preferences and satisfaction. Information garnered from text analytics helps identify critical variables and characteristics that can increase sales or marketability of a product, company, or service, or increase customers' positive attitudes toward a brand. For these reasons, CI practitioners are increasingly turning to text analytics techniques to derive insights from the vast amounts of textual data previously unavailable to them.

TEXT ANALYTICS DEFINED

We define text analytics as a collection of methodological techniques designed to explore, investigate, and examine attitudes, thoughts, patterns, and opinions found in text. Approaches to text analytics include content analysis (analysis of text using conceptual constructs that are based on theory or empirically driven scientific findings) or text mining (analysis of text without the use of a priori conceptual structures). Regardless of the approach, text analytics generally follows certain procedural steps, including data collection (or text identification), data coding, data analysis, and interpretation (see Figure 1).

Data Collection

For content analysis studies, data collection is the process of identifying relevant text sources and retrieving or formatting text for coding to create structured databases. Some companies already have websites in place that gather comments from customers. Public forums such as blog sites or topical web forums also generate independently gathered text. Other more established data sources for content analysis include emails, open-ended questions on surveys, and focus groups.

Alternatively, text mining can generate useful information from call center logs, customer feedback emails (customer suggestions or complaints), faxes, letters, employee suggestions, and customer relationship management (CRM) databases. You can also apply the technique to secondary text, such as financial filings, press releases, and articles, as well as high-profile speeches. In many cases content analysis and text mining are not mutually exclusive; rather, utilizing both can serve as an excellent validation technique.

Text Analysis

After obtaining text information, researchers employ both content analysis and text mining software to extract patterns and meaning from the information. Specifically, text analytics software does the following:

- Identifies synonyms.
- Groups related words.
- Extracts discussion themes.
- Extracts opinions/sentiment.
- Explores term patterns and term relationships.

Many software packages specialize in various types of text analysis, such as psychological trait analysis or semantic cluster analysis.

Text Interpretation

Most text analysis results in structured data. You can then apply quantitative techniques to the structured data to predict and model consumer behavior. Such techniques include cluster analysis, multivariate regression, logistic regression, or modeling techniques such as structural equation modeling or path analysis.

CONDUCTING TEXT ANALYTIC INVESTIGATIONS

Text analytics can be both an exploratory process and an empirical investigation. In content analysis, you begin with a priori hypotheses or preconceived ideas of what words or themes are related to certain characteristics or thoughts. The process is theory driven. In text mining, you work without preconceptions and base your findings on information "from the ground up." This process is data driven rather than theory driven.

Content Analysis (Searching for Emotion/Motive)

In content analysis, you use theory and experimental methods to identify the themes, words, and phrases that most commonly express a characteristic, emotion, or idea. In many cases, you can induce a particular emotion or state in respondents or identify people who exhibit a certain trait. You may even obtain writing samples from those respondents and respondents who serve as comparison or control groups. Writing samples can be obtained via primary research: responses to questions, stories in responses to pictures, opinion pieces about a given topic, or even respondents' attempts to complete an unfinished paragraph.

Various normative data sets are available for comparison. They may include internal company databases, data sets developed by text analytics software companies, or data sets from academia. Comparing two groups of writing samples extracts markers that indicate specific traits, emotions, or ideas.

Certain emotions and characteristics are strongly related to specific words. When people feel a particular emotion or believe in a particular concept, they are likely to choose specific words when expressing opinions. For example, when people are pleased with a product or situation, they generally use positive adjectives when giving their opinion. Psychologists also surmise that underlying subconscious thoughts, goals, motives, and traits also impact the words people choose.

Once the markers have been identified, you can translate your findings in word-phrase dictionary coding schemes. The program then scans the text to identify whether these dictionary words or phrases appear. Higher levels of a particular trait become apparent. The greater the number of dictionary words appearing in a text, the higher the level of a particular trait or feeling.

Example: Content Analyzing Respondent Text

Content analysis is commonly used to examine open-ended questions on survey data. Applying content analysis here helps discern subtle or subconscious attitudes or thoughts that might increase our understanding of respondents' behaviors. In one client study, respondents received a survey on vacation and travel. The study intended to identify traits and characteristics that might predict whether or not people would buy a vacation property. Among specific questions on demographics, travel preference, and recent vacation activities, the survey included this open-ended question:

> *Imagine you are offered an opportunity to buy a vacation property in a location you frequently visit. Would you buy this property? Why or why not?*

Responses ranged from the rare one-word answer (yes or no) to lengthy paragraphs discussing the pros and cons of buying a vacation property. Most answers contained one paragraph stating the decision and one to three reasons for the decision. Content analytical schemes specializing in measuring emotions and salient attitudes provided profiles of potential buyers and non-buyers.

Example: Content Analyzing Small Groups

Content analysis also examines recorded interviews and translates them into text. A specific example is a study that uses focus group discussions to determine how a new product might affect potential consumers. The structure of a focus group not only measures individual responses, but also uses them to reflect group dynamics and sentiment. With a focus group you can cover several topics in great depth. But its design is completely dependent upon prepared and scripted formats that reflect your agenda.

Content analysis can use the transcripts of focus group studies and discern subtle patterns and themes that may have not been specifically discussed by the group or are not apparent to the moderators. For example, in a recent Anderson Analytics study, when focus group members discussed a new beverage, they commented positively and voiced opinions that could classify them as future consumers of the beverage.

The application of content analytical investigation showed distinct differences among the group. Some group members emerged as discussion leaders and their profiles differed distinctly from the others. The leader's comments included many positive adjectives and emotionality. In contrast, the comments of other individuals were less positive

and contained instances of inhibition, uncertainty, and anxiety. The subtle differences detected by content analysis helped to elucidate the difference between people who were more likely to be positive and act on those emotions versus yea-sayers who just went along with the group but may not necessarily be future consumers.

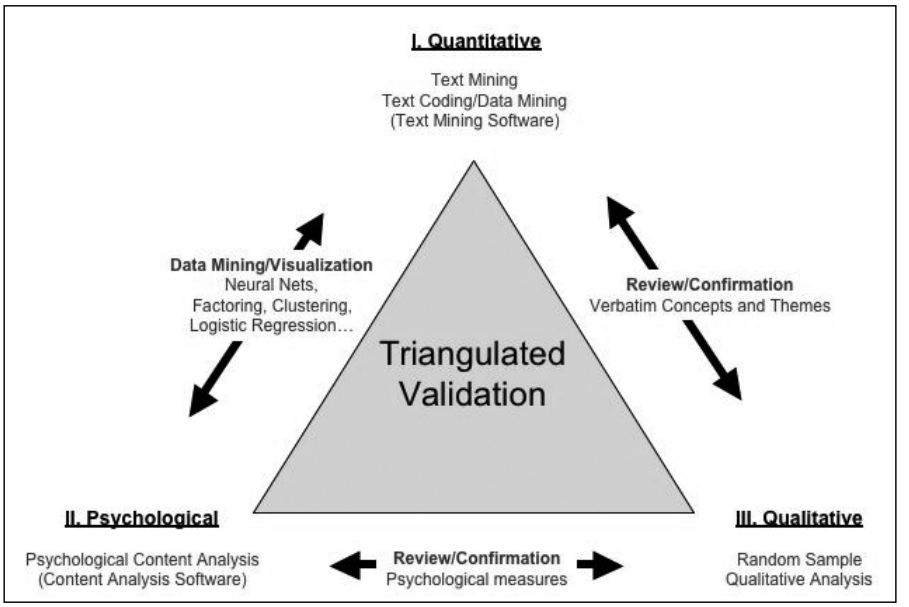

Figure 2: Anderson Analytics Pyramid Model for Text Analytics℠

Text Mining (Searching for Explicit Meaning and Sentiment)

In contrast to content analysis, text mining analysis involves the use of computer software and human coders to extract patterns and information, regardless of whether theories or empirical works guide the pattern identification. The software initially detects patterns, themes, and categories encompassed within the text. These analyses usually focus on word or phrase frequency, the appearance of other terms in conjunction with these words or phrases, and the conceptual distance between these terms (for example, the degree of relationship between words). This theme acquisition is not driven by a priori hypotheses or theory; the patterns are data driven and may not make sense on their own.

Human coders discern patterns by placing the identified themes into context to derive their meaning. For example, software first identifies that "soda" is frequently used in posts on a soft drink company's website. The program then discerns that when posters mention "soda," the adjectives "delicious," "fizzy," and "love" also frequently appear. You can then surmise that the four words "soda," "delicious," "fizzy," and "love" are conceptually related for some of the respondents who post on the company's website.

Example: Large Structured Data Sets

In certain instances, analyzable text has previously been extracted, formatted, and organized into structured data sets. These data sets become excellent sources for analyzing text for patterns and themes that help you understand consumer and, in other contexts, employee behavior.

Many companies keep customer service call log records to examine and maintain customer relationship quality control. These call logs contain information regarding the calling customer (name, date, gender, location, type of customer, topic of issue) and the agent who received the call (agent ID, years of employment, previous record of complaints, location, department). In many cases, they also contain transcripts of the conversation itself. This becomes an excellent opportunity to extract information regarding the interaction between callers and agents.

For example, you can use text mining to extract common complaints voiced by customers, the severity of their complaint, and whether the issue was resolved by the end of the conversation. The text database can also examine the strategies individual agents use to address customers who might have problems or complaints (for example, dismissive, compliant, confrontational, apathetic, or empathetic). You can then determine whether certain strategies work better for different types of problems. By focusing on the interaction between customer and agent, you can examine advantageous combinations and use them to help train and improve customer service within the company, thereby reducing customer churn and increasing the up-sell.

Example: Large Unstructured Data Sets and Web-Scraping

One way of exploring customer sentiment is to administer an opinion survey and document the responses to multiple-choice questions. This method can produce quick responses, but survey respondents may not readily provide their true attitudes or opinions. They may resent taking a survey and haphazardly answer the questions. Others may just ignore the request.

Still others may feel the need to respond congenially, replying in ways they think they are supposed to respond, but that do not necessarily reflect their own true opinions. In this case, try comparing these findings with an information source that is completely unsolicited and naturally occurring, such as text from websites or discussion boards.

Even though website data may more accurately represent consumer opinion, it is "noisy" and disorganized, as is free-form information on the internet. In this case,

you need to extract usable text from websites using "web-scraping" or "screen-scraping" software that scours selected websites for specific data of interest. These programs then extract information such as text, date of post, and poster identification such as forum ID or screen names. After extracting the text and accompanying information, standard text mining procedures, including software coding, can discern variables based on word frequency and proximity of these words to other key words. Human coding then provides context and meaning to those variables.

For example, in an Anderson Analytics study on the hospitality industry conducted on flyertalk.com earlier this year, the first set of categories generated by a text mining program included "nights" and "stay." Human coders who understood the nature of the discussion (the actual exchange between the primary actors) instructed the program to group those two terms together. Using their knowledge about the industry and the information guideline obtained from the initial concept extraction, the coders continued to modify the categories until they obtained satisfactory levels of detail and category logic.

TEXT ANALYTICS PYRAMID/TRIANGULATION MODEL

As demonstrated above, the differences between content analysis and text mining are centered in their approach to data analysis. Content analysis uses a top-down approach, applying theory or empirical evidence to coding methods. Text mining uses a bottom-up approach, extracting patterns from text and using human coders to interpret patterns and discern meaning. But these approaches are not necessarily at odds with one another. They can complement each other, as demonstrated by the Anderson Analytics Pyramid/Triangulation Model of Text Analysis (see Figure 2).

By coupling content analysis and text mining with qualitative readings of the text, you can support, validate, and create layers of information within the analyses. In the triangulation model of text analysis, the initial analysis phase requires conducting all three forms of analysis in parallel so that initial findings are independent of one another.

The text mining portion of the analysis identifies relevant groups or variables within the data. The content analysis phase measures the general states and emotions of the respondent sample and can be compared to other normative samples such as national data sets or norms based on empirical data. The qualitative analysis identifies context and themes indicative of the topic or sample. Later process phases involve cross-analysis—using information garnered in the initial stages to help understand and elucidate findings in individual analyses.

You can use variables identified in the text mining as grouping variables to recognize differences in content analysis results. Text mining results might classify a sample into three groups: potential consumers, potential distractors (those that take our attention away from something in terms of value, importance, or quality), and neutral groups. You can use information garnered through content analysis to profile these groups by their emotions, attitudes, and characteristics. For example, potential distractors are low in happiness, high in anxiety, and high in stubbornness. Potential consumers are high in happiness, low in inhibition, and high in agreeableness. Neutral respondents are high in inhibition, uncertainty, and concern for money.

Qualitative findings can confirm these profiles, or amend them based on an in-depth reading of the text. For example, potential consumers are high in happiness and low in inhibition, as evidenced by this quotation: "I love this product. It makes me feel warm and fuzzy inside. I would not hesitate to go out and buy this product."

THE FUTURE OF TEXT ANALYTICS

In many situations, text analytics can help analysts better understand their customers and even their own organization. Each situation calls for a particular type of text analytics. Even so, the diversity of text analytics allows easy application to the various needs seen in market industries today.

Consumers, businesses, and competitive intelligence practitioners will see an increase in the use of text analytics. Text analytics has not only been applied more often, but also appears more frequently in scholarly research. Academics and researchers from the fields of computer programming, linguistics, psychology, sociology, bioinformatics, medicine, and engineering have discovered the utility in applying text analytics to their own investigative pursuits.

You can look forward to additional advances in text analytics and its application to obtain actionable information about your consumers and your competition. Text analytics holds the promise of being the richest yet most cost-effective source of intelligence. It may also be the best way to first identify new trends (opportunities or problems) before they become widely apparent.

Tom H.C. Anderson is managing partner of Anderson Analytics LLC, a next generation marketing consultancy that specializes in combining new technologies, such as text mining, with traditional marketing research. He has worked on product and market development projects in multiple countries and industries and is a frequent lecturer in graduate-level marketing research and data-mining courses. Tom is the author of "Listen to the text" published in Quirk's Marketing Research Review, October 2006, and has presented on text mining at conferences including SPSS Directions, ESOMAR Automotive, ESOMAR Leisure, and the USA Text Analytics Summit 2007. He can be reached at toma@andersonanalytics.com.