Paper SP06-2011

What Are People Saying about Your Company, Your Products, or Your Brand?

Kathy Lange, SAS Institute Inc., Cary, NC

Saratendu Sethi, SAS Institute Inc., Cambridge, MA

ABSTRACT

A growing area of interest for many firms is understanding what the market (customers, analysts, or key opinion leaders) is saying about their products and services. Many refer to this area of analysis as sentiment analysis. They want to understand more about people's opinions, attitudes, and emotions when discussing their products, services, or overall brand.

From the company perspective, listening and analyzing what people are saying about products and services is the first step in creating a dialogue with that audience by listening, learning, and then engaging with them. This dialogue can better inform targeted marketing initiatives to customers and prospects, enabling the organization to communicate at a significantly lower cost than traditional marketing with increased speed and effectiveness. It can also enable a more rapid response to perceived customer issues and competitive threats.

This paper discusses the use of SAS® Sentiment Analysis and SAS® Text Miner to uncover good and bad feedback. It discusses lessons learned from real projects.

WHY DO ORGANIZATIONS CARE?

All types of businesses and government agencies are starting to take advantage of information buried in previously untapped text documents (surveys, product reviews, online forums, e-mails, instant messages, articles, etc.). They are interested in protecting their brand equity, increasing customer satisfaction and loyalty, and reducing risks by carefully guarding their reputation. Initiatives like these are often driven through the sales and marketing departments, but are monitored by top executives including the CFO, COO, and CEO, as they have far-reaching financial impact.

Top executives attempt to manage all types of risk within their organizations. Many of these risks are financial (credit risk, market risk, liquidity risk, insurance risk), but others are nonfinancial. Reputational risk is one of the highest nonfinancial risks identified in the "Global Risk Management Survey: Sixth Edition" executed by Deloitte in 2009 [1]. That survey identified that 81% of the respondents were attempting to manage reputational risk. Firms need to understand what the market (customers, analysts, or key opinion leaders) is saying about them, their products, and their services, and prepare for any unforeseen issues. The following are some examples of companies (or people), where public incidents have caused an impact on their reputation. The actions that are taken after an incident occurs are often the key to whether they recover from damage incurred to their reputation.

- Tylenol (product tampering)
- Domino's Pizza (employee video)
- Wendy's (finger in the chili hoax)
- Toyota (sudden acceleration reports)
- Ford/Firestone (tire issue)
- Citigroup (purchase of corporate jet following government bailout)
- President Clinton (Monica Lewinsky scandal)
- Marion Jones (runner stripped of Olympic medals for performance-enhancing drug use)
- Mark McGwire (Major League Baseball steroid scandal)
- Potential employees (impact of negative items posted on Facebook)

MANAGING REPUTATION

Reputation can take decades to build, but can be lost very quickly. A strong reputation can be a key competitive advantage. Creating an ongoing dialogue with customers or stakeholders to listen and learn from their feedback, in good and bad times, can help establish and maintain a good reputation. This dialogue can better inform targeted marketing initiatives to customers and prospects, enabling the organization to communicate at a significantly lower cost than traditional marketing with increased speed and effectiveness. It can also enable a more rapid response to perceived customer issues and competitive threats.

Companies interested in managing their reputation are often looking for answers to the following questions: What are the issues that people are discussing about my business compared to the top competitors in my industry? Are my marketing messages resonating with my customers? Am I spending my marketing dollars on the right channel to reach my target audience? What actions should be taken based on the feedback I am hearing? What effect will the feedback have on my brand and my reputation in the marketplace? The following sections look at the process of assessing the answers to some of these questions.

SENTIMENT ANALYSIS: THE PROCESS OF ASSESSING AND SCORING OPINIONS

Sentiment is an attitude, thought, or judgment prompted by feeling (Merriam-Webster). ***Sentiment analysis*** is the process of determining and measuring the tone, attitude, opinion, and emotional state of responses. More precisely, it is the concept of deciding whether a specific conversation is positive, negative, or neutral. Figure 1 demonstrates an example of sentiment identification in an online review site.

I really like the design of the phone and the strength of the signal but the call center was slow and the rep was impatient.

- Assign *positive* sentiment to **phone features** (design, phone) and **network coverage** (signal strength)
- Assign *negative* sentiment to **customer service** (call center)

Figure 1. Sentiment Analysis Example

Sentiment analysis has broad applications and encompasses work in classifying subjectivity, polarity, tonality, emotion mining, opinion mining, persuasion analysis, and affective computing. It is a tool that allows companies to analyze what their customers are saying regarding their products and services, and also monitor trends in the opinions and attitudes of their customers toward the products and services with respect to their competitors.

APPLICATIONS OF SENTIMENT ANALYSIS

Traditionally, companies had to rely on marketing campaigns to understand sentiment. With the advent of Internet and online social media, various forms of online expression have replaced person-to-person social interaction. Companies are increasingly interested in methods of online sentiment summarization (also referred to as social media monitoring). The measure of sentiment trends serves as a virtual currency for businesses to strategize their public relations, marketing, and sales efforts. Among the many sectors that are analyzing text to uncover sentiment are retailers (including grocery), banks, insurance companies, airlines, hospitality, online services, telecommunications, technology sectors, and government. Aspects of the customer experience are often evaluated for positive and negative sentiment.

The following are a few industry examples that describe applications of sentiment analysis:

- For airlines, the topics of interest might be in-flight experience (legroom, food, and entertainment), ticketing, baggage claim, price (fees), routes, and schedules.
- For insurance companies, the topics of interest might be overall satisfaction with the company, cost, coverage, service, and claims processing.
- Clothing retailers might be interested in fit, style, comfort, availability, and customer service.
- Pharmaceutical companies might be interested in overall brand sentiment, promotions, customer service, price, documentation, and ordering. They also might be interested in segmenting the responses from consumers, health care professionals, industry advocates, and key opinion leaders. There is rising interest in understanding the sentiment of drugs that have recently gone generic (off-patent). They might want to analyze sentiment on social media and blog sites to determine where they would get the most lift from placing promotions or coupons for these “mature brands.”
- Financial institutions are considering examining all employees’ e-mails and instant messages to detect inappropriate discussion of financial instruments. This action would identify whether they are being discussed in a positive or negative way, which might be a security breach or an indication of insider trading.
- Government agencies, now mandated to incorporate voter feedback, might be interested in feedback on laws, policies, proposed bills, or school assignment plans.

- In the political arena, questions might include: What is the sentiment around political candidates and the issues they are discussing? What is the ideology of politicians or political groups based on what they are saying or publishing? Are they really liberal or conservative? Can we modify the candidate's stated political positions to gain acceptance by major voting segments?

AUTOMATED SENTIMENT ANALYSIS: NEED AND CHALLENGES

It is impractical for humans to find, read, evaluate, summarize, and organize the sentiment of the vast number of articles, blogs, and social media conversations. Companies have to increasingly rely on automatic methods for sentiment analysis.

Locating various opinion sites and monitoring them on an ongoing basis can be a formidable task due to the scale of the Internet and the volume of content on each site that is produced daily. In many cases, opinions are hidden in long forum posts and blogs. Extracting sentiment from text is a challenging problem with applications throughout natural language processing and information retrieval. Below is a list of the primary tasks involved in implementing an automated sentiment analysis system along with the challenges associated with them:

- Data acquisition
  - Identifying data sources
  - Data quality issues (cleansing, preparation, processing, scalability)
- Sentiment analysis through natural language processing
  - Community lingo, including slang (e.g., WTF, OMG), exclamations (e.g., !!!, ?!), emoticons (e.g., ☺, ☹)
  - Word ambiguity (e.g., Amazon the online retailer vs. river vs. tribe in Greek mythology)
  - Clarity of Intent, including sarcasm (e.g., "hell" vs. "hell yeah!"), context (e.g., "I don't like the store but I get great bargains there!")
  - Language constructs, including parsing (part-of-speech, stemming), negated Idioms and expressions (e.g., anything but, hard act to follow, not bad), entity extraction (e.g., product names, brands)
- Scope of sentiment analysis
  - Define topics of interest and identify emerging topics
  - Granularity of sentiment analysis
- Delivery of results
  - Visualization (e.g., dashboards, charts, reports)
  - Operational issues (e.g., hosting vs. in-house)
  - Downstream activities (e.g., response mechanisms, campaign strategy, customer relationship management)

For the rest of the paper, we describe each of the phases in detail and present how SAS products help you solve these problems.

DATA ACQUISITION

Data acquisition is the first step in sentiment analysis. It involves source discovery, content aggregation, filtering, and metadata tagging. The result of data acquisition is a feed of relevant content data, with duplicates, spam, and other noise removed, which can be processed for sentiment and mined for specific information. Relevant, real-time, quality data allow businesses to make more timely, accurate, and informed decisions to gain significant competitive advantage.

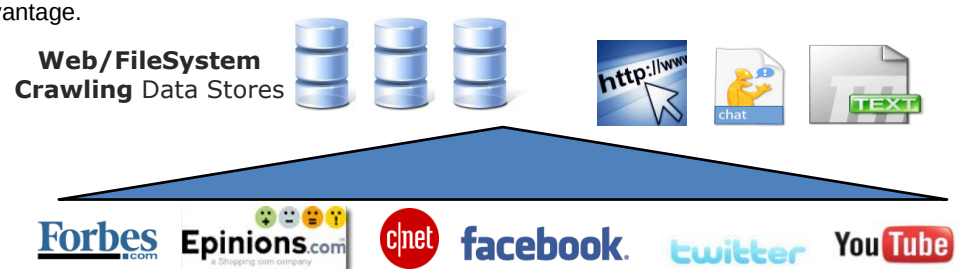


Figure 2. Content Sources

Data acquisition needs to strike a fine balance between downloading the universe and gathering content that is too specific. Relevant data for companies is spread across a variety of sources, including online sources like blog posts, discussion boards, social networks, and newsgroups; enterprise systems like intranet, customer relationship management (CRM) data, enterprise e-mails, and private message boards; traditional media sources like press releases and news articles; and multimedia sources like podcasts, radio, and television programs. The sources can be proprietary (i.e., require access through subscriptions, web services, and custom APIs) or open to all (e.g., online sources).

While setting up the process for data acquisition, a key aspect to consider is the need for constraining the data explosion. Popular social media sites like Twitter and Facebook have been growing at an astounding rate of 1444% per year [2] and 500-million-plus users [3], respectively. International Data Corporation (IDC) has predicted, in a study sponspred by EMC, that the world enterprise digital data would reach 1.2 zettabytes (10^{21} bytes) in 2010 [4]. Hence the queries for data acquisition must be constrained to obtain only the relevant data for processing without biasing the content.

The discussion of various proprietary online sources is excluded from this paper due to the specific nature of their APIs and restrictions (e.g., rate limits, licensing costs). The rest of the Internet and intranet data can be downloaded by using a scalable crawling tool like SAS® Web Crawler [5].

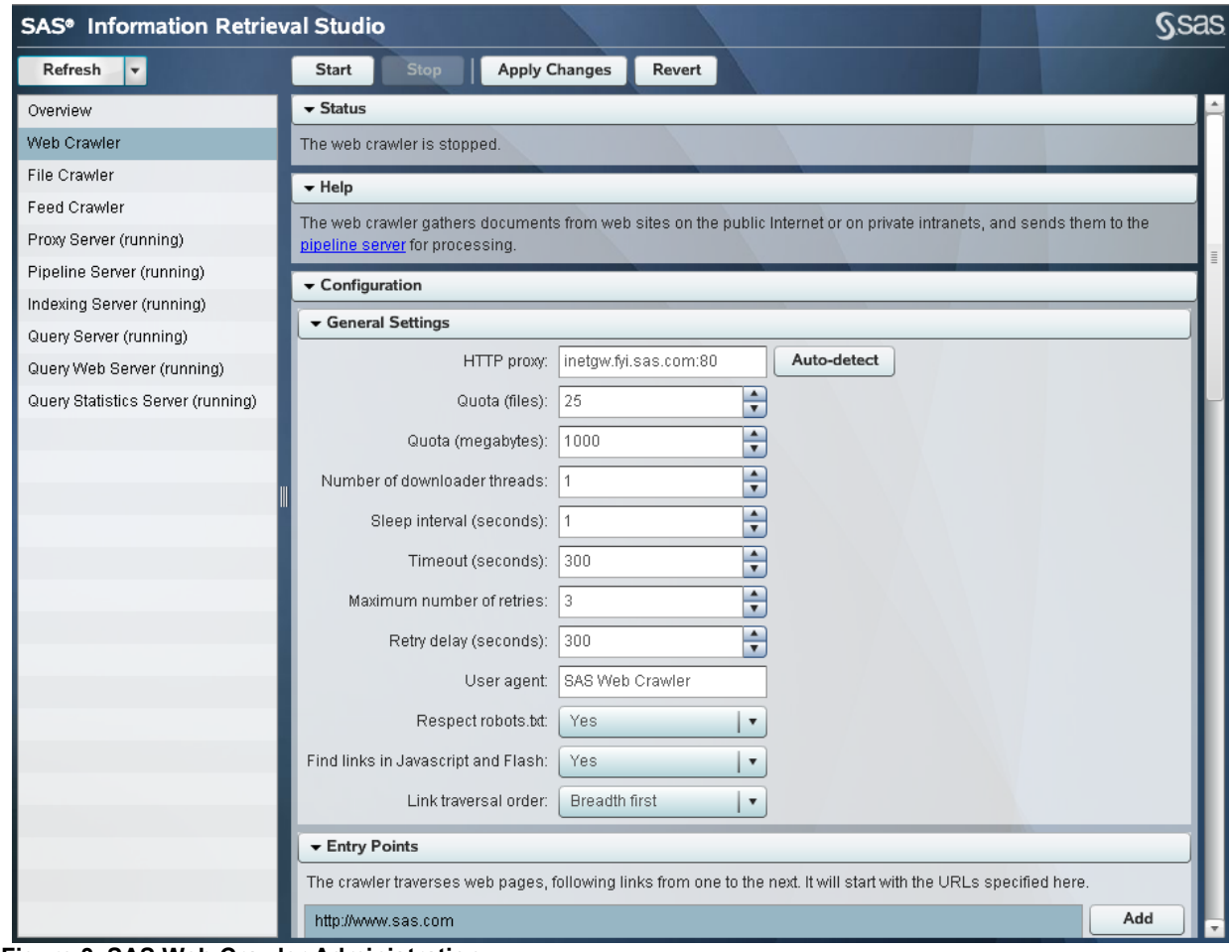


Figure 3. SAS Web Crawler Administration

The crawler is a sophisticated program that automatically downloads documents from the Internet. It is a multi-threaded program that allows companies to define starting URLs of interesting Web sites or internal file shares. Parameters can easily be defined for controlling the behavior of link traversal through those sites.

For scalability at the level of the Internet, the crawler should support distributed processing and methods for intelligent link selection through methods like language identification and automatic categorization to download useful content only. For example, the crawler should be able to interface with automatic categorization to filter out unwanted content, such as job postings, spams, duplicates, and advertisements, while following links through pages with relevant

content about the company, brand, and products. We describe more about automatic categorization in further detail in the next section.

After the content is downloaded, the data typically undergoes various cleansing processes and is stored in an enterprise data warehouse (EDW) or data mart in a normalized form for further processing.

SENTIMENT ANALYSIS THROUGH NATURAL LANGUAGE PROCESSING

The research community has approached sentiment analysis through a variety of techniques from the areas of text mining, natural language processing, and computational linguistics. For most data (particularly with evolving lingos in the online community), it hasn't been possible for automated sentiment analysis technologies to perform on-par with a well-trained human indexer. Albeit, automated methods still remain the preferred approach for their scalability, speed, consistency, and ability to reprocess data with updated sentiment models. The accuracy of automated methods has long been debated given the subjective nature of human sentiment attribution. Human beings, typically, do not exhibit 100% agreement with each other (it has been claimed to vary between 60–80% [6]) due to their personal level of knowledge, their influence from surroundings, and their tacit motivations.

Typical text mining approaches have involved machine learning techniques like Bayesian inference [7], latent semantic analysis [8], and support vector machines [9]. Each of these methods works from a vector-based representation (e.g., term-document matrix), and hence are generally referred to as "bag of words" approaches. The major drawback with these approaches has been the inability to maintain context information in the term-document vectors. As a result, the performance of text mining approaches is severely restrained.

Natural language processing (NLP) approaches, on the other hand, are able to capture the specificity of sentiment nuances through a variety of concepts, like part-of-speech disambiguation, sentence parsing, entity extraction, and context-based Boolean operators. They usually involve a form of rule-writing, which requires up-front manual labor but, once developed, enables the desired accuracy levels within the subjective context. Tools like SAS® Sentiment Analysis [10] use a hybrid approach that benefit from the best of both worlds (Figure 4).

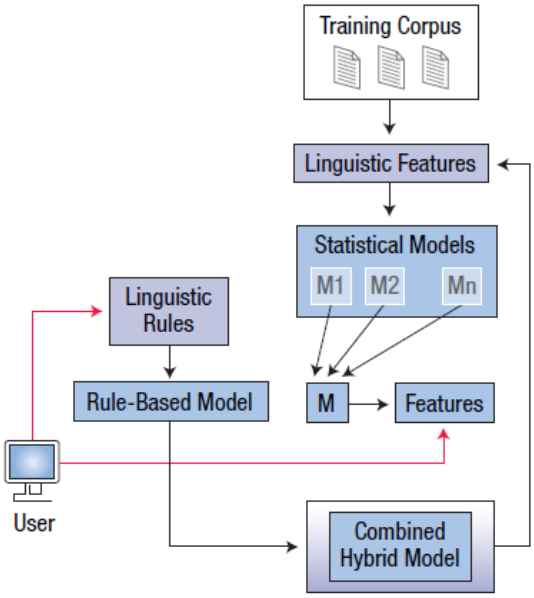


Figure 4. Hybrid Sentiment Analysis

In this section we talk about some of the primary NLP approaches to solve specific sentiment analysis challenges. The basic idea is to approach sentiment analysis as a lexicon-building exercise that involves building dictionaries of positive-negative terms. Depending on the specificity of the domain, these terms can be enclosed within contextual boundaries to express certain situations. For example, one might consider "cheap" as positive in relation to price, but negative in relation to quality. Below we provide additional examples of various types of terms used in this approach.

Figure 5. Natural Language Processing Techniques

a) Community-speak terms: Involves specific symbols and terms that are clear indicators of sentiment:

"***I love my bank as the employees are very helpful*** ☺" → ☺ → positive sentiment
"***Are you kidding me****?!*" →?! → negative sentiment

b) Polarity terms: Involves building lists of sentiment words and phrases:

"***terrific***", "***awesome***", "***the best***" → positive sentiment
"***hell***", "***awful***", "***the worst***", "***waste of money***" → negative sentiment

c) Clarity of intent: Sometimes we need to add some context (e.g., handle sarcasm):

"***hell yeah!***" → positive sentiment
(**NOTWITHIN**, "***hell***", "***hell yeah!***") → Uses Boolean operator (NOTWITHIN) to clarify that "hell" expresses negative sentiment only if it doesn't occur within the expression "hell yeah!" (which is a positive exclamation).

Various NLP tools provide different sets of Boolean operators that can broadly be classified into the following classes:

1) Boolean operators: AND, OR, NOT
2) Counting operators: Number of occurrences and count of distinct terms
3) Proximity operators: Operators based on word distance, scope of sentence, paragraph
4) Contextual operators: Operators based on order, XML fields, position within the document, variety of NOT operators (e.g., NOTWITHIN).

d) Language constructs: Involves the use of part-of-speech tagging, entity extraction, and concept-references:

1) Negated concepts (e.g., sequences like "***not :Adjective***", "***not :Adverb :Adjective***"), which usually convey negative sentiment like "*not good*", "*not going well.*"
2) Entity extraction: Extract product names and attributes:
"<***sequence of words***>™": A sequence of words in the context of ™ in unstructured text could be identified as a product name, which can be the focus of sentiment attribution. SAS® Enterprise Content Categorization and SAS Sentiment Analysis provide a powerful feature that identifies all references to a product name throughout the document if the product name has been identified at least once in context (see more specific examples below).
3) Concept references: Use intermediate concepts in Boolean rules to identify sentiment: **(SENTENCE, (*WORD_DISTANCE_5, "def{PRODUCT}", "def{POSITIVE_WORDS"}*))** where the overall rule is looking for positive sentiment by finding product names in the context of positive words. ***def{PRODUCT}*** and ***def{POSITIVE_WORDS}*** here are references to rule sets for identifying product names and positive sentiment words that are defined elsewhere.

4) Pronoun resolution: Allows the linking of specific function words to nouns: **(SENTENCE, *"pronres{ :Pronoun }", "PERSON"*)**, which expresses that all mentions of pronouns as identified by part-of-speech tagger should be assigned to the person mentioned in the same sentence.
5) Co-referencing: Allows writing rules to identify and link two entities as synonyms: **(SENTENCE, *"coreference{def{PRODUCT}}", "def{COMPANY}"*)**, which expresses that product names should be used as synonymous to those company names when they are mentioned at least once together in the same sentence.

SCOPE OF SENTIMENT ANALYSIS

The success of a sentiment analysis deployment can be measured only if it is helpful in analyzing the content according to the business needs. The results should be able to provide insights and tangible cost savings to the end users. Looking back to the "Applications of Sentiment Analysis" section of this paper, it can be surmised that organizations evaluate a taxonomy of "high value" aspects of their business. For example, a retail bank might be more interested in learning the sentiment around banking aspects rather than the sentiment around political and social issues. A banking taxonomy would help to restrict the scope of sentiment analysis to the banking context.

A taxonomy is primarily characterized by the structure and information represented in its hierarchical organization when compared against forms of controlled vocabularies, like lists and synonyms. Taxonomy classification is a method of classifying and assigning structural information to unstructured data. Taxonomies can be of various types:

- Functional taxonomies: Represent the business model or specific function or service (e.g., sentiment analysis)
- Topics taxonomies: Categorize and organize data by topics or nature of the content (e.g., content categorization)
- Organization taxonomies: Represent departmental functions (e.g., customer service classification, marketing functions)
- Directories and site maps: Organize information for Web-based content (e.g., enterprise information management)
- The work of well-known researchers in sentiment analysis [11, 12] further corroborates the need to use taxonomies to assign sentiment at a granular object-feature level. SAS® Text Analytics provides multiple methods of taxonomy management through SAS Sentiment Analysis and SAS Enterprise Content Categorization. (See Figure 7. Taxonomy Management.)

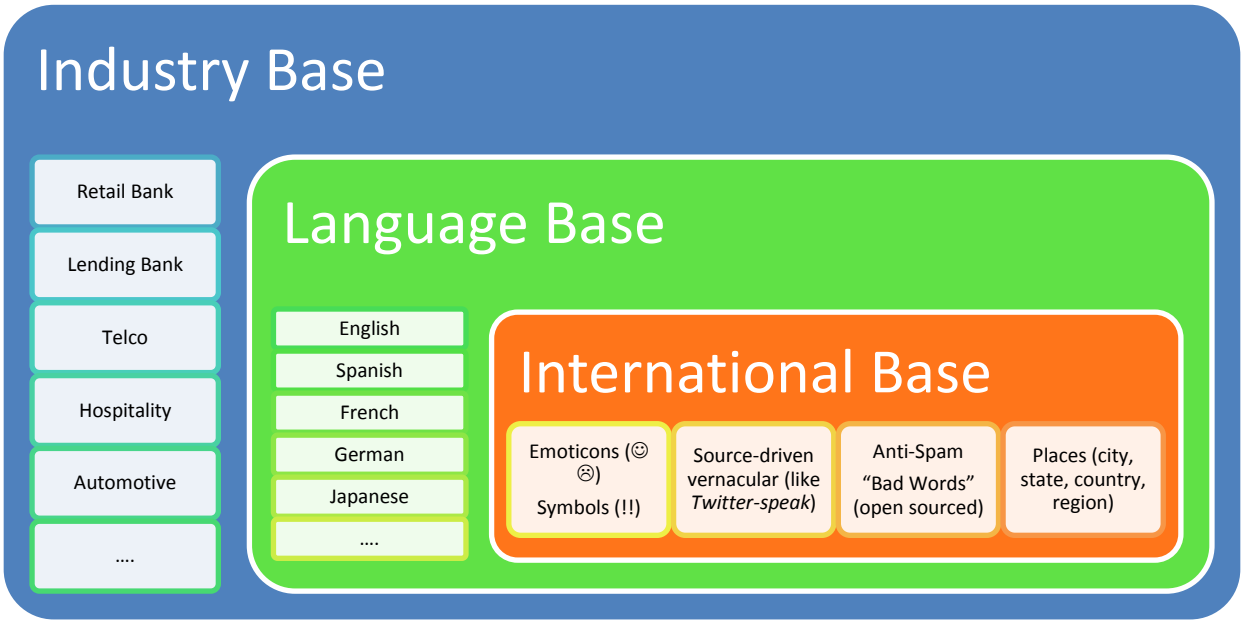


Figure 6. Business Taxonomy Defines a Scope Around Sentiment Analysis

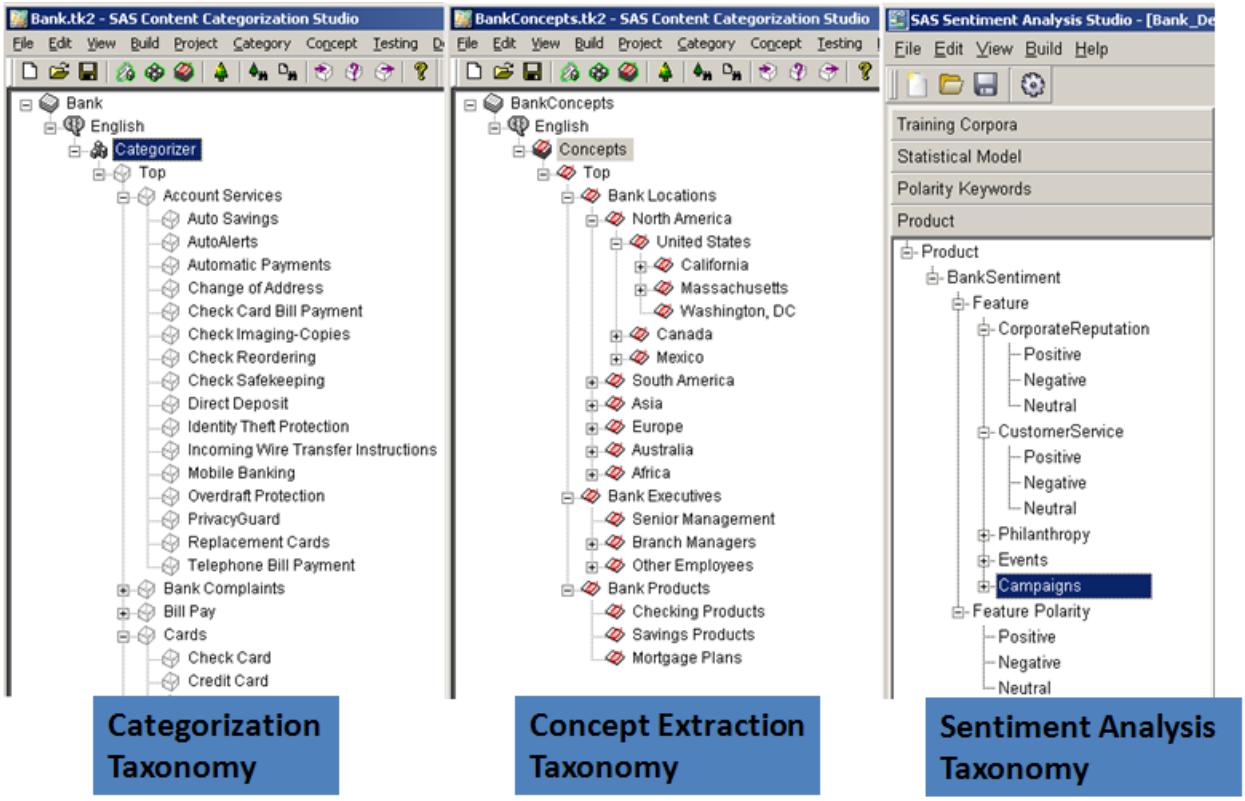


Figure 7. Taxonomy Management

DELIVERY OF RESULTS

The results of sentiment analysis need to be validated and summarized by business analysts to make conclusions. The analysts need to sieve through tremendous amounts of data coming through online resources, press releases, and news publications.

They need access to a workbench or dashboard interface that enables them to monitor and search for actionable insights using a set of custom analytics tools. One such tool is SAS® Sentiment Analysis Workbench [5], which allows analysts to collaboratively define corpora of documents, validate and modify the results on any document, execute ad hoc search queries, and generate customizable charts, summaries, and reports to understand and measure consumer feedback and differentiate their products and brands. The workbench interface also provides a feedback channel for correcting misclassifications and to progressively migrate from a purely manual solution to a scalable and automated solution without any loss of precision and control.

Figure 8. Analyst Interface

A SENTIMENT ANALYSIS USE CASE

In a recent analysis of product reviews for various brands of printers obtained from several merchant Web sites, we looked at specific product features, services, and customer interaction points with the manufacturer. The taxonomy for this analysis was:

- Overall company or brand
- Service (service, support, salesperson, staff)
- Ink (ink jets, cartridges, refills, ink tank, ink life)
- Cost (price, payments, dollars, bucks)
- Printer models (system, model name, series)
- Other general attributes (installation, instructions, ordering, Web site)

Statistical methods and text mining tools in SAS Sentiment Analysis and SAS® Text Miner were used to explore the documents and to "learn" domain-specific terms and phrases that were associated with positive and negative reviews (relative to these printer reviews). Stems of words and synonyms were automatically identified as well as misspellings of the words. Synonyms and entities were also automatically extracted from the text in the text parsing phase.

Closely related terms and concepts were identified through link analysis, and various themes and topics emerged through document clustering. Descriptive terms of each of the clusters are seen in Figure 9, in the output of SAS Text Miner. These terms are displayed for each cluster beginning with the highest-weighted terms.

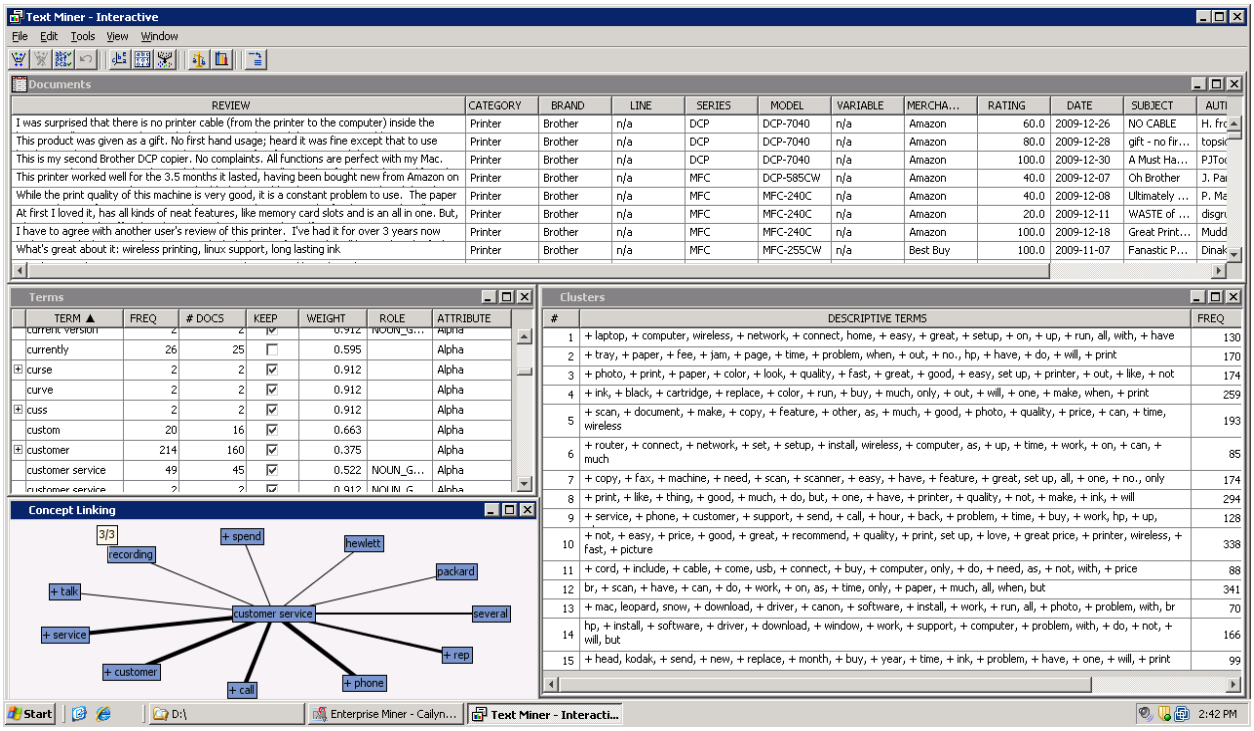


Figure 9. SAS Text Miner Exploration

Linguistic rules were developed using the insight gained from the statistical modeling and text mining to identify positive and negative sentiment associated with each of the taxonomy topics, creating a hybrid sentiment model (combining both statistical models and rules-based models). In this analysis, all the rules for each manufacturer were identical (except for specific product names and models), in order to compare each of the "features" across manufacturers consistently.

Figure 10 shows reports generated from the analysis for the ink and service topics in the taxonomy. The bars represent the percentage of positive and negative comments, rather than the absolute number, to normalize for the number of reviews for each of the manufacturers. This enables us to compare the brands more easily.

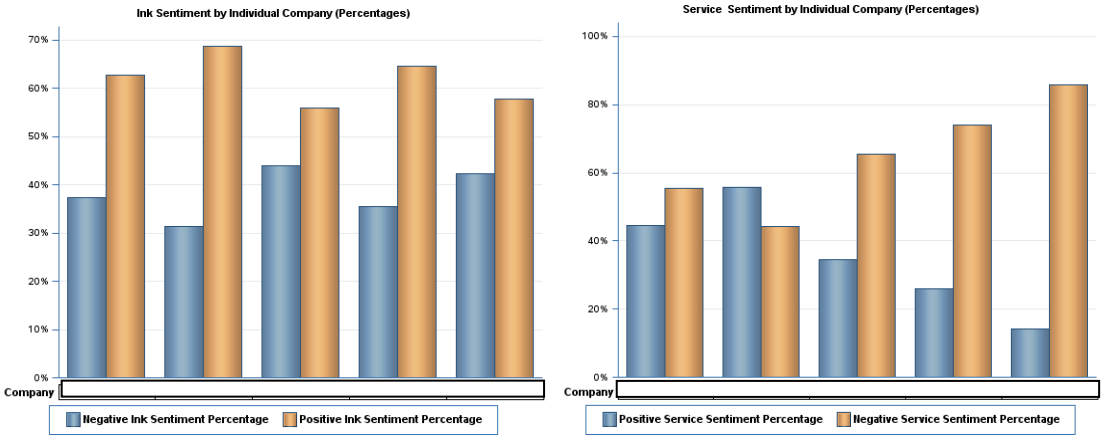


Figure 10. Sentiment Analysis Output (Ink and Service Categories) by Brand

Other themes discovered in the text mining exploration included envelope-printing issues and USB cable issues (which you might expect), but we also discovered a “Christmas gift” topic.

Along with the actual text in the review, the SAS Web Crawler retrieved structured information that was associated with each review from the merchant’s Web site. This information included a numerical rating, printer model, location, date of review, and so on.

By combining the text with structured data, we were able to perform profiling of the various clusters/segments as well as predictive modeling using both the structured and unstructured data in addition to new variables that were created by the SAS Sentiment Analysis “features."

In the segment profiling analysis (Figure 11) we identified that one merchant, Best Buy, was more highly associated with positive reviews in the segment dealing with “scanning” (segment 12). Graphical tools in SAS® Enterprise Miner were used to visualize the profiles of each merchant in this segment as compared to the merchant reviews of the overall document corpus.

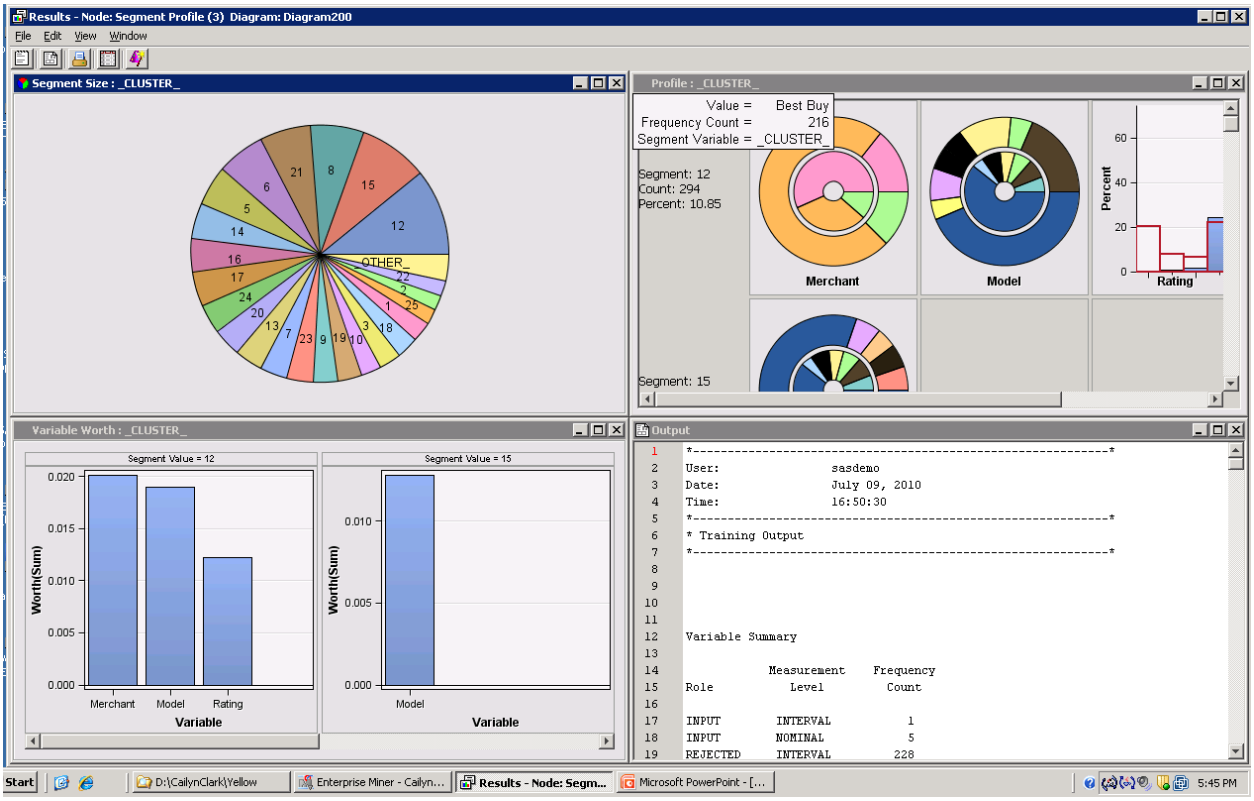


Figure 11. Segment Profiling

Predictive models were developed to predict overall product ratings in the hold-out sample. Combining the text cluster information as an additional variable in the analysis yielded significantly better results for every model (decision tree, regression, neural nets) that included the clustering information. In addition, the new variables derived from SAS Sentiment Analysis “features” turned out to be highly predictive in the decision tree, adding more “lift” to the models (Figure 12).

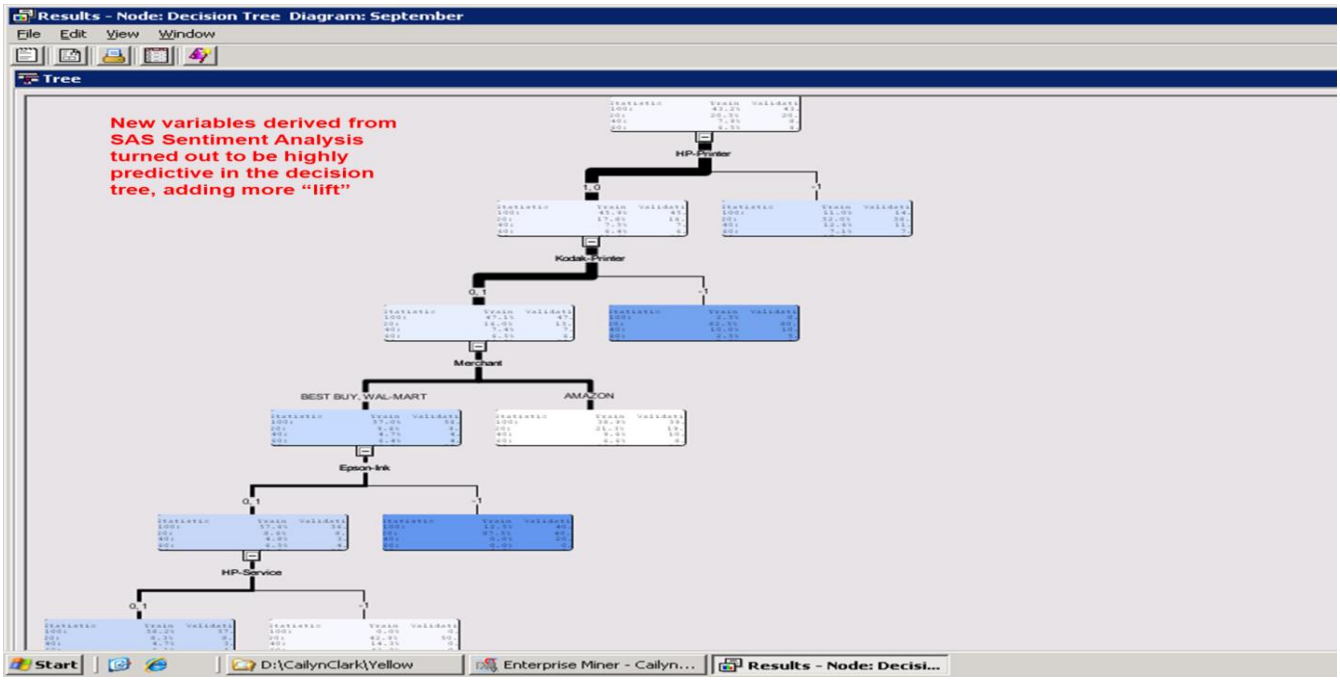
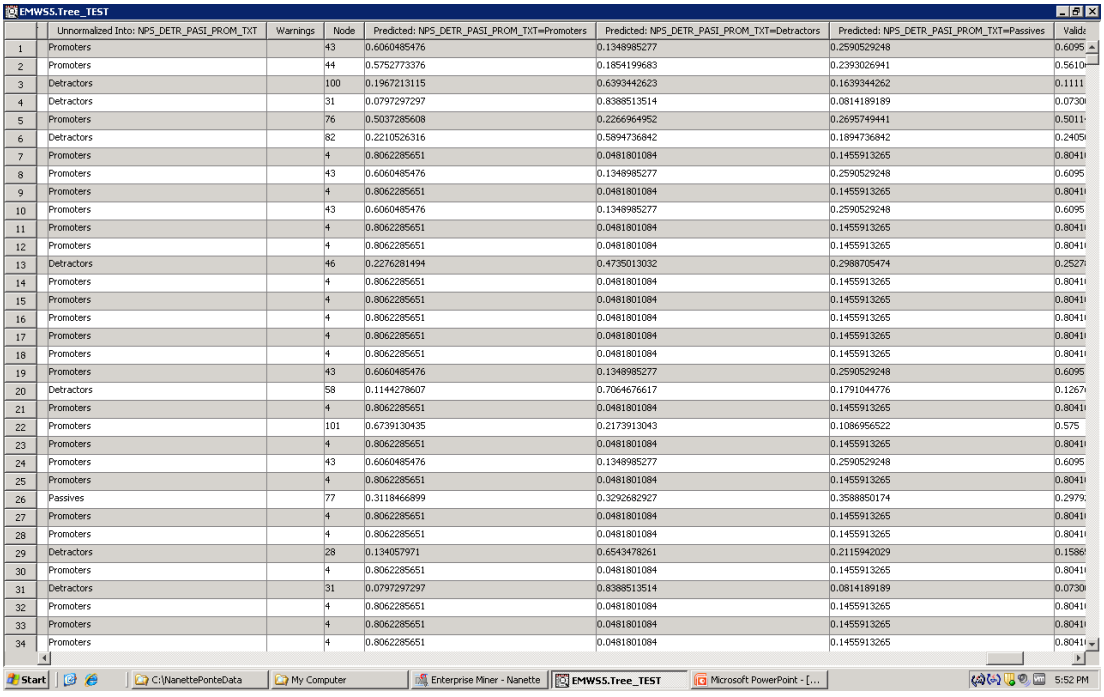


Figure 12. Decision Tree Output with New Sentiment Analysis Variables

More and more often the rating that companies are trying to predict is Net Promoter scores. The sentiment around a particular aspect of the customer experience can be a strong predictor of customer loyalty (predicting whether a customer will be a promoter, detractor, or passive). The text mining analysis can derive probabilities for each of those categories. (Figure 13)


EMWS5.Tree_TEST

	Unnormalized Into: NPS_DETR_PASI_PROM_TXT	Warnings	Node	Predicted: NPS_DETR_PASI_PROM_TXT=Promoters	Predicted: NPS_DETR_PASI_PROM_TXT=Detractors	Predicted: NPS_DETR_PASI_PROM_TXT=Passives	Valid
1	Promoters		43	0.6060485476	0.1348985277	0.2590529248	0.6095
2	Promoters		44	0.5752773376	0.1854199683	0.2393026941	0.5610
3	Detractors		100	0.1967213115	0.6393442623	0.1639344262	0.1111
4	Detractors		31	0.0797297297	0.8388513514	0.0814189189	0.0730
5	Promoters		76	0.5037285608	0.2266964952	0.2695749441	0.5011
6	Detractors		82	0.2210526316	0.5894736842	0.1894736842	0.2405
7	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
8	Promoters		43	0.6060485476	0.1348985277	0.2590529248	0.6095
9	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
10	Promoters		43	0.6060485476	0.1348985277	0.2590529248	0.6095
11	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
12	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
13	Detractors		46	0.2276281494	0.4735013032	0.2988705474	0.2527
14	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
15	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
16	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
17	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
18	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
19	Promoters		43	0.6060485476	0.1348985277	0.2590529248	0.6095
20	Detractors		58	0.1144278607	0.7064676617	0.1791044776	0.1267
21	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
22	Promoters		101	0.6739130435	0.2173913043	0.1086956522	0.575
23	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
24	Promoters		43	0.6060485476	0.1348985277	0.2590529248	0.6095
25	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
26	Passives		77	0.3118466899	0.3292682927	0.3588850174	0.2979
27	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
28	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
29	Detractors		28	0.134057971	0.6543478261	0.2115942029	0.1586
30	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
31	Detractors		31	0.0797297297	0.8388513514	0.0814189189	0.0730
32	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
33	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041
34	Promoters		4	0.8062285651	0.0481801084	0.1455913265	0.8041

Figure 13. Probabilities That Each Review Would Be Associated with a Promoter, Detractor, or Passive

DOWNSTREAM BUSINESS ANALYTICS

Business analytics is about making better decisions and is a continuous improvement process. There are a variety of people within an organization that are involved in the analytically driven decision making process that analyzes customer feedback (Figure 14).

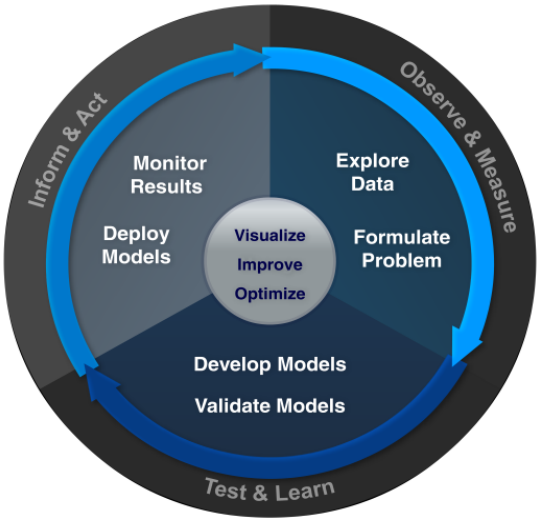


Figure 14. Business Analytics

Information builders will be tasked with building the proper reports, dashboards, or spreadsheets that can be automatically populated with new information as it is being processed. *Power users* or *analysts* will likely work with subject-matter experts (or industry-domain specialists) to build the statistical and data mining models as well as the linguistic rules that categorize topics discussed (observe & measure and test & learn in Figure 14).

Models, taxonomies, and business rules need to be continually updated and monitored over time to incorporate new or changing information. New words might come into the vocabulary over time. They could be new technology, slang terms, colloquialisms, or mashups of existing words. For example, “mahusive”is a mashup of “massively” and “huge”. Is “sick” or “wicked” positive or negative in regard to the topic being discussed? New terms might find their way into existing topics. For example, in the National Highway Traffic Safety Administration (NHTSA) data, there has long been a topic around “sudden acceleration” reported in the data. Until recently, that topic was associated with a wide range of automobile companies. Of late, Toyota has become much more prevalent in association with that topic.

New topics can emerge that were not previously identified. Early detection of new issues is critical, particularly if they involve safety issues. But other issues can arise that could affect customer satisfaction and loyalty.

In many of our customer’s data, we see dissatisfaction in the customer service topic of the taxonomy around the service agents that do not speak English well. It is a prevalent problem with large call centers, especially those that are outsourced overseas. Recently, we also noticed another topic that identified negative sentiment around the customer service agents reading from what sounded like a script, or sounding like a robot, or having canned responses. This issue should be monitored over time to see if there is a growing negative trend or whether it is consistent and steady.

Sales and marketing departments, product management, and top executives want to understand what consumers are saying, but have little desire to be involved in statistical modeling or in developing linguistic rules to distill the mounds of text that are being collected. They are merely “information consumers”, who will most likely be tasked to take appropriate actions based on the insights that have been discovered in the feedback (inform & act in FIgure 14).

The following are examples of insights that need to be evaluated for potential actions:

- Customers view your product return policy negatively.
- Customers view your price more negatively than your competitors, even though you market your products as the low-cost provider.
- Nurses are very positive about a specific drug that you manufacture (in comparison to doctors, consumers, other medical personnel).

- There is significant negative sentiment about customer service representatives that have non-English accents.
- High-value customers are growing more dissatisfied about fees associated with in-room Internet services in your hotels.
- Several states are more highly negative about customer service issues than others (Figure 15).

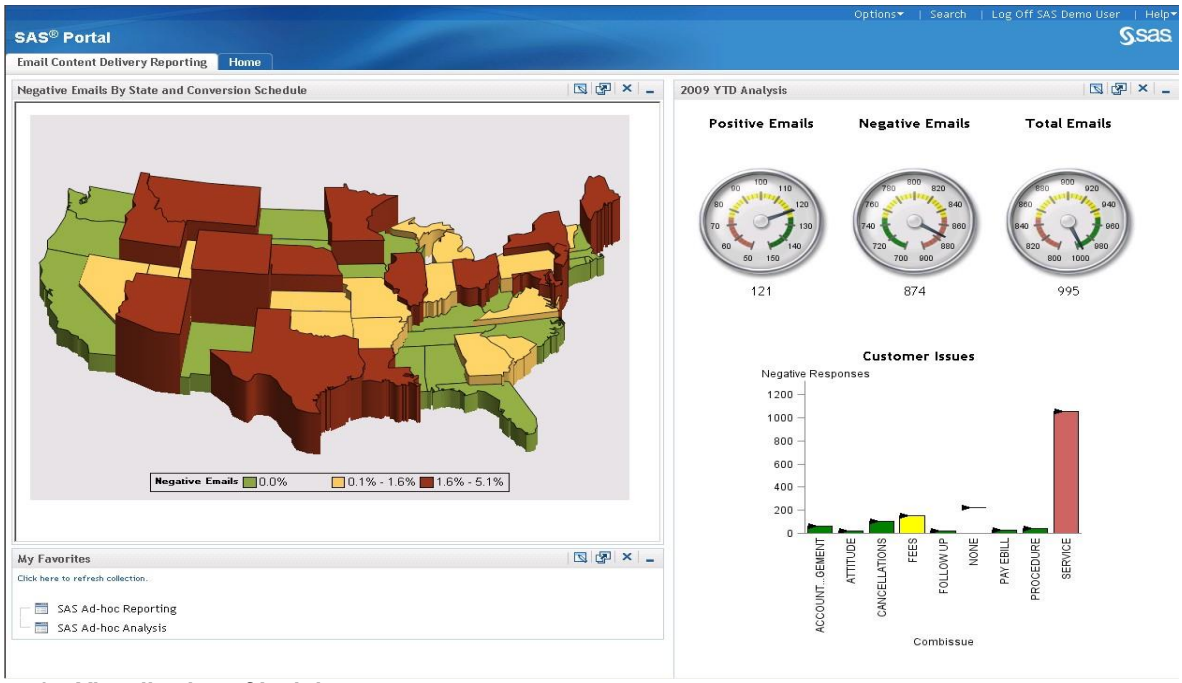


Figure 15. Visualization of Insight

The insights might lead you to identify additional training needs of staff, product or process improvements, or marketing opportunities to a specific customer segment. This is the reason that organizations need to evaluate "high value" aspects of their business, in order to respond to the feedback and take actions that provide measurable return to the business.

CONCLUSION

Text is a largely unused asset in many organizations. Firms need to interpret, summarize, and report on information that is contained in documents. They understand the virtues of moving beyond reporting to proactive business analytics that are forward looking to reduce uncertainty, predict with precision, optimize performance, and minimize risks.

The same processes and methods that are used for analyzing structured data can also be applied to text. Analysis of large corpuses of text can be explored and analyzed using automated tools, rather than having humans read each document. Text can be used to augment the analysis of structured data to gain more insight and greater lift.

Sentiment analysis can use both a statistical and a natural language rule-based approach. The major advantage of a rule-based method is the amount of control it gives the rule developers over how the analysis will be performed. Developers can use their knowledge of the domain and the specific language within it to develop rules that have high precision. The results of rule-based analysis are easily interpreted and provide full transparency into exactly what sentences, keywords, or context within the document triggered the positive or negative sentiment.

REFERENCES

1. "Global Risk Management Survey: Sixth Edition." Deloitte. Available at http://www.deloitte.com/view/en_US/us/Industries/Banking-Securities-Financial-Services/center-for-financial-services/7f18773b93912210VgnVCM100000ba42f00aRCRD.htm.
2. Nielsen News. June 22, 2009. "Twitter Grows 1,444% Over Last Year; Time on Site Up 175%." Available at http://blog.nielsen.com/nielsenwire/nielsen-news/twitter-grows-1444-over-last-year-time-on-site-up-175/.
3. Mark Zuckerberg. "500 Million Stories." The Facebook Blog, July 21, 2010. Available at http://blog.facebook.com/blog.php?post=409753352130.
4. John F. Gantz, et al. March, 2008. "The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011". Framingham, MA: IDC. Available at http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.
5. SAS Institute Inc. SAS Text Analytics. Available at http://www.sas.com/text-analytics/.
6. Theresa Wilson, Janyce Wiebe, and Paul Hoffman. 2005. “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis.” In *HLT '05: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354. Stroudsburg, PA: Association for Computational Linguistics.
7. "Bayesian Inference." Available at http://en.wikipedia.org/wiki/Bayesian_inference.
8. Latent Semantic Analysis. http://lsa.colorado.edu/.
9. Corinna Cortes and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning* 20:273–297. Available at http://www.springerlink.com/content/k238jx04hm87j80g/.
10. SAS Institute Inc. SAS Sentiment Analysis. Available at http://www.sas.com/text-analytics/sentiment-analysis/index.html.
11. Bo Pang and Lillian Lee. 2008. “Opinion Mining and Sentiment Analysis.” *Foundations and Trends in Information Retrieval* 2(1-2):1–135.
12. Bing Liu. 2010. "Sentiment Analysis and Subjectivity." In *Handbook of Natural Language Processing*, 2d ed., Edited by Nitin Indurkhya and Fred J. Damerau. Boca Raton, FL: CRC Press.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Kathy Lange
Enterprise: SAS Institute Inc
Address: 100 SAS Campus Drive
City, State ZIP: Cary, NC 27513
Work Phone: (919) 531-4192
Fax: (919) 677-8000
E-mail: kathy.lange@sas.com
Web: www.sas.com

Name: Saratendu Sethi
Enterprise: SAS Institute Inc
Address: 10 Fawcett St. Suite 6
City, State ZIP: Cambridge, MA 02138-1175
Work Phone: (617) 576-6800 Ext: 54246
Fax: (617) 576-6888
E-mail: Saratendu.sethi@sas.com
Web: www.sas.com