Seeing the Voice of the Customer: Identifying Root Cause with Text Analysis Visualization
Table of Contents

Introduction ......................................................... 1
Visual Text Analytics Best Practices .......................... 2
Visualizing the Customer Voice .................................. 5
Root Cause Exploration ............................................ 6
The Sound of Sentiment ........................................... 9
Identifying Relevant Points of Interest ......................... 10
Case Study: Super Bowl XLVII Blackout ...................... 11
Conclusion .......................................................... 14
For More Information ............................................... 15

Content for this paper was initially provided by Mary Osborne, Justin Plumley and Daniel Zaratsian. Osborne is a Systems Engineer Manager for SAS Global Sales Development and Product Management. Plumley and Zaratsian are Technical Consultants for SAS Global Sales Development and Product Management.
Introduction

Stop to think about how – and how often – your business interacts with customers. Every day, with each interaction, data is created. What percentage of the data generated by these interactions are you using? There are so many channels for interaction, like social media, call centers, sales staff, help and support resources, marketing and campaigns. Typically, organizations believe that they are using only a small fraction of it effectively – at best, upwards of 10 percent of all the available data. Why?

One reason relates to the difficulties in collecting all this data. This limitation is beginning to wane as commodity hardware becomes increasingly popular for big data storage. But another major inhibitor to examining all customer data has been the inability to examine millions, or even billions, of data points that constitute the customer picture. And much of this is now in the form of unstructured text inputs.

Unstructured data continues to grow in volume, variety, velocity and the overall value it provides to organizations. Today, customers are voicing their preferences, problems, concerns, issues and accolades in the form of customer feedback, survey responses, social media conversations, blogs, documents, maintenance notes, news articles and more. At the same time, there is a continuing shift in the ways to derive value from this unstructured data.

There are many ways to understand text data. Depending on your experience, background or business requirements, you may begin to evaluate your text data with a simple word cloud that illustrates the emphasis of words and phrases that occur within text collections. Many methods – topic identification, text clustering, natural language processing, sentiment specifications and taxonomies – are used today to understand the key themes, categories, concepts and entities present in text documents.

Each of these methods provides some sort of structure as a natural output. Whether they come in the form of term counts, topic membership identifiers, sentiment scores or variables, they can be used to classify and organize text materials. These structures can also be used to establish more complex relationships between text content by using ontologies and network graphs. No matter where you fall in this spectrum, your objective is the same: to extract meaningful and usable information from raw text data. Often, this requires interactive and exploratory queries.

Due to recent technological innovations, the roadblocks to examining all customer data have been removed. Former sampling requirements and filters that once limited customer activity investigation to 10 percent or less have dissipated. Analysis is no longer restricted by long batch processing times. IT departments are no longer required to predefine data models for interactive insights. Users no longer need to make requests for specific data prior to queries. Now, all your customer data can be available for you to explore without any constraints being presupposed, defined or assumed.
One enabling solution is SAS® Visual Analytics. This software liberates companies to examine all of their customer data, allowing stakeholders, analysts and data scientists to derive customer insights as quickly as they can think of the next question to ask. The results of text analysis provide an initial structure to unstructured content – and this easy-to-use, visual technology designed for big data allows you to dynamically interact, explore, investigate and query the text data to quickly investigate relationships, patterns and outliers.

By integrating SAS Visual Analytics with SAS Text Analytics, organizations are empowered to uncover patterns and relationships within both structured and unstructured data. This combination creates an enhanced view of your big data. It enriches and provides a visual look at your data that reveals customer sentiment, shows categorical flags and uncovers root causes that primarily exist in unstructured data.

Text analysis and the subsequent visualization are naturally symbiotic. Because even the best analysis of unstructured data is often of little value unless you can readily communicate the results. Easy-to-understand visualizations explain any analysis, whether your audience includes colleagues, executives, customers or investors. Of course, the data that drives good visualizations needs to be accurate, correctly formatted, properly constructed and relevant in terms of aligning with your business requirements. Text analysis provides this relevancy.

**Visual Text Analytics Best Practices**

Business requirements and questions define the rationale for any analysis – regardless of whether the inputs to that analysis are traditional structured data, unstructured data or a combination of both. Just as with extraction from existing structured customer data stores, text data is retrieved from social media sites, customer agent transaction notes, email communications, forums and the like. Figure 1 illustrates that information retrieval is the first step in analytical processing.

*Figure 1: Visual text analysis process flow diagram.*
In many cases, organizations have both internal and external data – structured and unstructured – that they want to combine and use for visual analysis. For example, in the financial services industry a banking customer may be affiliated with call center notes from financial advisers, transactional data based on banking activities, demographic data from applications, and questions or commentary about the bank that they posted in social media channels.\(^1\) Note that particularly with big data, the goal is not to physically move and integrate different sources of customer data – but to affiliate the data in such a way as to create a single, virtual and federated record that references each customer.\(^2\)

The unique challenge with text data is that it is free-form. This opens the door for misspellings, individual styles, abbreviations, emoticons, misused terms (e.g., writing “their” instead of “they’re”), and many other data quality challenges. In addition, the analyst typically doesn’t know what the raw text collection contains to begin with – so assessing the documents with text mining becomes an essential analysis step in this process, as depicted in the second step of Figure 1. Text mining explores and extracts key elements in the data, such as the relationships between terms and phrases; and it generates topics and clusters in an automated or semiautomated, machine-driven process. Text mining results in statistically derived part-of-speech analysis and standardizations, and this includes the automated identification of term stems, synonyms and misspellings.\(^3\)

The next step, dependent upon business requirements, is to enrich the original data set by creating new variables that identify sentiment and categorize the unstructured data. This is depicted in Figure 1, steps 3a and 3b. This step may be the most time-consuming because it requires the analyst to define taxonomies and rules that extract the desired elements from the text data. Advances in SAS Text Analytics reduce this burden, with the automatic generation of Boolean rules that can directly inform linguistic definitions for taxonomies. Initial training data sets for sentiment analysis can also be mined from the text collection using the AFINN data set included with SAS Text Miner.

Step 3 is the most modeling-intensive stage, but it’s also one of the most valuable. In this step, you produce extracted and tagged unstructured data, which turns free-form text into structured data.

---

\(^1\) SAS has the ability to collect and aggregate inputs from disparate structured and unstructured sources of customer information with SAS/ACCESS® and the SAS Crawler (for both internal file systems/intranets and external websites, including RSS feeds and common social media sites like Twitter, Facebook, Google, Bing, YouTube).

\(^2\) Federated views span multiple data repositories, enabling data from different systems without the physical reconciliation or movement of source data, and achieved for analytical tasks with SAS/ACCESS technology and/or the SAS Federation Server.

\(^3\) Additional information on text mining operations is available at: [sas.com/text-analytics/text-miner/index.html](http://sas.com/text-analytics/text-miner/index.html)
Regardless of your organizational goal – to enhance predictive models with unstructured data, organize and categorize content, assess sentiment or accomplish a combination of any of these – you often need to take a dual approach to model development. This involves using statistics and machine learning to uncover what you don’t know about the text content (step 2), as well as taking a linguistic rule-based, taxonomy approach (steps 3a and 3b), to accurately incorporate domain experts’ knowledge with the automatically discovered patterns. This approach of using both discovery methods and domain-driven results generally provides the greatest flexibility to appropriately model text data.

Step 4 of Figure 1 shows the enriched structured data, which includes the extracted and tagged fields derived from the text analysis. As the diagram illustrates, the data is now formatted for use in other activities, like those listed in step 5.

For predictive analysis, insights derived from text data can improve the predictive power of existing models (i.e., increasing model lift) and in some cases can even result in a better model than those developed from structured data. These outputs also create metadata that can improve existing retrieval systems – like enterprise search – surfacing more relevant material based on the additional insight that references the document content. They can even surface semantically related materials. These structured outputs can be used to trigger threshold alerts, with notifications sent to downstream operational applications or stakeholders who can be notified of priorities when they are highlighted from analysis of the customer documents.

Finally, these structured text documents can be used in visualization and reporting to help you explore, understand and uncover the driving factors affecting customer activity and impression. At this stage, it’s important to know whether the goal of the visualization is analytical exploration or operational reporting.

If your goal is data exploration, it may be beneficial to include variables that allow for general pattern discovery. For operational reporting, you may already know an existing scope. For example, with exploration you’d want access to the enriched data set that contains not only the original data of interest, but also new flags, hierarchies and derived variables to gain further insight into any identified patterns. Alternatively, if operational reporting is your goal, the specific key metrics that you need to communicate to business stakeholders wouldn’t require an augmented data set with additional fields, but would simply be contained in the file and represented in the reports.

---

4 Predictive model power can be compared across models, and those developed solely on text data may outperform existing models based solely on structured data as illustrated in From Customer Risk to Corporate Strategy: Using Text Analytics and Predictive Modeling to Improve Promoter Scores: sas.com/reg/wp/corp/59208
Visualizing the Customer Voice

You can visually explore unstructured customer data once the text sentences and phrases have been tokenized and parsed. This identifies the elements within the text (usually words, expressions, punctuation marks, white space, stems, etc.) and analyzes the grammar.\(^5\)

During the initial text mining step, you can visually explore the terms and phrases associated with the customer communication collection to help guide the mining process. Using interactive discovery with this vast data can help pinpoint key themes within the complete text collection and also narrow the focus for your analysis to desired concepts and filter out noise in the data.

Figure 2 illustrates concept linking,\(^6\) which illustrates associations and relationships between terms. The more highly associated the terms, the thicker the line between them.

![Concept Link Illustration](image)

**Figure 2: Concept link illustrating term associations for the word “print.”**

From Figure 2, we can see that the term “print” is more strongly associated with “large” and “small,” as well as with “light,” “print quality” and other aspects of the type or purchase experience.\(^7\) You can drill into each of these term boxes to understand associated terms and phrases from the collection. By discovering these relationships, you can get insight as to how terms are used together within the document collection – and this can guide the creation of a taxonomy.

---

\(^5\) Tokenization, parsing (the grammatical analysis of a sentence/phrase) and stemming (variants of words like pluralization, conjugation, case, etc.) are all automatically performed within the Text Parsing node of SAS Text Miner or SAS Contextual Analysis. The user has configuration controls to override system defaults.

\(^6\) Concept linking is generated from the Interactive Filter Viewer of SAS Text Miner, or the term map of SAS Contextual Analysis.

\(^7\) The “+” sign prior to terms listed in the concept-linking diagram refer to additional stems that the system found associated with that specific word/phrase.


**Root Cause Exploration**

In the context of document publishing, if the term “print” is associated with the terms “large” and “small,” it might make sense to create a parent node in a taxonomy called “print” with a child node called “small.” This approach would enable you to appropriately classify the key factors and their variants. Taxonomies are simply a hierarchical set of categories (and often subcategories) that describe the relationships among terms or phrases within a document collection. For this data, a simple taxonomy may look like Figure 3.

![Figure 3: An example taxonomy.](image)

Once the taxonomy has been developed, you can use it to classify the customer document collection. In fact, contextual extraction can be used to identify when a term may be used as a noun or as a verb – to provide even deeper insight to what issues may be occurring. For example, “print” (as a noun) or “font” can be extracted in the context of the same sentence as the word “small” (including synonyms) or “large” (including synonyms). These terms, as well as others associated with “print,” can be used as the foundation for a root cause analysis.

Continuing with our publishing dilemma, we can see that there are issues related to “print.” And having defined our taxonomy, scoring our customer communications with our linguistic model, we can now explore and interrogate our structured text data with interactive visual analytics. As Figure 4 shows, it’s easy to identify the driving factors relevant to “print.” We can see that “print” and “font” occur in sentences with the term “small” (and its synonyms) far more often than with any of the other terms in the hierarchy.

---

8 SAS Enterprise Content Categorization provides taxonomy development and linguistic rule support to create contextual extraction models that are highly configurable to the document collection.
And if we continue to drill into the underlying factors associated with “small,” we identify a strong relationship between terms that mean or indicate “print” and words that describe “print” typeface (in relation to terms customers use to describe issues they are experiencing).
Drilling one final time into the more frequently identified term reveals sentence snippets that provide the actual context in which the terms were used. This is depicted in Figure 6.

![Figure 6: Hierarchy drill-down into the terms “small” and “font” to discover root cause.](image)

With a hierarchy such as this defined, we can quickly drill from high-level concepts into actual words of customers. We can see that the three predominant factors are, in the words of customers: the “font is so small,” the “font is very small” and “small the font.” These are the underlying issues associated with small print/font typeface. Clearly, the legibility issue is due to a typeface that is overtly small to meet the needs of customers. By knowing that this is the underlying issue, you can immediately recognize what remedial action is necessary to address the problem.

This example highlights one of the core benefits of visually analyzing unstructured customer data. Not only are the most prominent issues across the entire collection of customer feedback readily identified – you can also derive the real context of what is driving the issue simply by drilling into the data. As is common with any text analysis, it’s the richness of insight contained in the commentary that explains why this is an issue. As a result, you can set strategies and tactics to rectify the situation.
The Sound of Sentiment

Sentiment analysis is often considered core to voice-of-the-customer projects. The ability to derive positive or negative tone from surveys, call center notes and social media can be critical to understanding customer reactions to products and services. The power of visualizing sentiment analysis results is that it enables you to examine the entire collection of feedback from customers, and to interactively scrutinize every individual element – so you can assess what is working and what requires corrective action.

In Figure 7, we see the results of the publishing data scored with a sentiment analysis.9 The red bars indicate negative sentiment while the green bars depict positive sentiment. We can see from the figure that “print” sentiment is more negative than positive, with the verbatim commentary associated with “print negative” listed in the table.

Figure 7: Sentiment analysis of “print-related” customer reviews.

This sentiment visualization illustrates for each (parent) category of the taxonomy that customers are very happy with the service they receive from the company, the overall quality and the prices being charged. They have equally mixed feelings regarding the shipping process and the offers they receive during that process – something to investigate further. The overriding negative feedback is associated with the print and order process.

---

9 SAS Sentiment Analysis can use existing taxonomies from SAS Enterprise Content Categorization, as well as user-defined taxonomies, linguistic rules and statistical models to identify and extract the tone associated with text inputs.
We can see from the detailed (negative) commentary what particular issues customers have experienced, and even suggestions from customers as to what could have been done differently. For example, the first comment was: “Your Web presentation of your card should be as close as possible to the printed card. Often the lettering on the card is smaller than the presentation to the point that not a few times the printing was too small to …” This is a viable suggestion: Include actual size representations on the Web as a way to adequately meet customers’ expectations.

SAS Sentiment Analysis taxonomies can be defined to any number of levels. So you can drill into each of these aspects of the business further to get detailed insight into the strengths identified by customer feedback as well as customer concerns. In short, this method exposes the root cause behind customer sentiment.

Identifying Relevant Points of Interest

Categorization and sentiment analysis both create structured information from the entire record of customer text input. For example, a record of customer communication could belong to a category like “print” or it could describe negative sentiment about “quality.” Sometimes, there may be a need to hone in on specific attributes of interest. These specific entities could describe a particular person (perhaps a known influencer), a place (different service locations), an event, or a fact (such as a particular product). For these situations, you can use entity extraction to identify the occurrence of these desired concepts. And once they’re identified, you can retrieve the associated content, filtering the unstructured information specifically associated with these items.

Consider social media data. By all estimates, it is growing exponentially. Some organizations are storing it in commodity hardware environments (like Hadoop) without first examining it to determine whether or not the data holds any potential value. Social data is big data, and to realize value from it, you must separate potentially valuable documents from irrelevant content and extraneous data.

Extraction can assist you, even with the initial storage of such data, by filtering out completely irrelevant content. For example, if you were the Bell company (a Canadian telecommunications corporation), you’d only be interested in consumer content relevant to your organization. You would not be interested in Bell Helmets, Bell Helicopter, Bell’s Brewery Inc., etc.

10 Entity extraction is included with SAS Enterprise Content Categorization, and is achieved with a defined taxonomy and configurable linguistic rules.

11 Refer to From Big to Meaningful Data for further description of unstructured big data issues and resolutions: sas.com/eg/wp/corp/59833
Visualizing taxonomies can also help to further delineate relevant from irrelevant content. Once scored with the taxonomy, text input contains structure that can be interactively queried with SAS Visual Analytics. Such visualizations help business users and subject-matter experts validate the classified results to determine if things classified as noise truly are irrelevant and should, consequently, be removed from the analysis.

On the other hand, perhaps this noise category contains new concepts not previously defined in the taxonomy – such as concepts arising within social media channels. You can use this insight to validate the existing taxonomy and hierarchies that represent the category/subcategory (or parent/child) relationships in the data. This can be built on the fly, helping you identify any new terms that need to be defined to the taxonomy.

**Case Study: Super Bowl XLVII Blackout**

Twitter users had a wide variety of things to say about Super Bowl XLVII, which was held in New Orleans on Feb. 3, 2013. The game was a matchup between the Baltimore Ravens and the San Francisco 49ers. More than 4.85 million tweets (from Jan. 11, 2013 through Feb. 6, 2013) were collected to see what fans had on their minds. Because of this broad timespan, the tweets covered not only Super Bowl XLVII discussions, but also Super Bowl XLVII predictions throughout the playoffs, content unrelated to football (such as the Beyoncé halftime performance), Super Bowl XLVII party chats, and a variety of speculation that occurred during the impromptu Super Bowl XLVII blackout.

Due to the wide variety of conversations and the nature of unstructured Twitter data, it took extensive data cleansing and filtering to better focus the analysis. For example, we needed to remove the irrelevant tweets, and identify misspellings and abbreviations. By identifying and grouping misspellings, synonyms and abbreviations with SAS Text Miner, we enriched the clustering results, eliminated unwanted noise in the data and aided in the discovery of emerging topics. The process was iterative and required some input from football aficionados. The result of this initial discovery phase resulted in industry (sports)- and organization (Super Bowl XLVII)-specific mapping of key terms and phrases that exposed insightful topics buried in the data. This is shown in Figure 8.

---

12 SAS offers a range of prebuilt industry taxonomies in different languages to quick-start the taxonomy development process. Additionally, initial taxonomy insight can be discovered from mining the document collection, and/or by automatically generating categories and subcategories from sources like Wikipedia or DBpedia.

13 For example, synonyms for the San Francisco 49ers included SF49, niners, 49ers, 49er, and more.
The term “power outage” was highlighted as the most prominent topic from the filtered Twitter data. Figure 8 also illustrates the calculated sentiment associated with the highlighted topic – which appears to be fairly balanced at this level of the analysis. Individual tweets related to the power outage topic are also shown.

One important aspect of evaluating the impact of social media is to assess the most influential contributors to the social media voice. In this case, “Influential Twitter users” were based on the number of followers they had, the frequency of tweets and re-tweets they made, and the overall frequency of Super Bowl XLVII-related tweets.14

One of the most informative ways to understand influence is from a network visualization analysis, which explicitly represents the relationships between topics and the most influential Twitter users. This is illustrated in Figure 9.15

14 Influence of any particular social media contributor can be based on a variety of factors, including calculated indexes. This case study used a simplified influencer definition.
15 There are a variety of ways to generate network graphs, including custom code or with SAS, using SAS/GRAPH®, SAS Customer Link Analytics, and SAS Social Network Analysis solutions.
From this visualization, we expose several significant conversations from Twitter users. During the second half of Super Bowl XLVII, which started at approximately 8:30 p.m. ET, the 49ers kickoff was returned by Jacoby Jones for a record-breaking 108-yard return, resulting in a Baltimore Ravens touchdown. On Twitter, it was interesting to discover some of the emerging topics at that same time. For example, one tweet stated: “This game is a joke.” Supporting tweets included phrases such as “this game is boring” and “game over!” – and there was even one tweet at 8:33 p.m. ET suggesting “Turn off the Lights #SuperBowl #NFL.”

This is a particularly interesting coincidence, because nearly five minutes later the lights did, in fact, go out. The blackout lasted nearly 35 minutes. During this time, a wide variety of conversations took place on Twitter. These topics, along with the authors who promoted the topics and the associations across topics, are all visualized in the network graph shown in Figure 9.
One very nimble marketing department – Walgreens – took advantage of the blackout by creating a social marketing promotion during that time. This promotion was identified within the data.

Although Walgreens\textsuperscript{16} did not come up with an ad visual, it only took seven minutes for this company to respond to the situation, by tweeting “We do carry candles.” This is shown in Figure 10.

![We do carry candles. #SuperBowl](https://via.placeholder.com/150)

Figure 10: Twitter conversation lead by Walgreens during Super Bowl XLVII blackout.

This tweet generated several re-tweets and engaged football fans to remember Walgreens when the lights go out or whenever they are in need.

Beyond the promotional social noise, we can learn even more if we continue to monitor and drill into the voice of the consumer by extending our analyses and visualizations to the article coverage that ensued post-Super Bowl XLVII. We could also track the impact on company sales from these low-cost and effective advertising tactics.

**Conclusion**

A growing number of organizations are realizing the value of unstructured data as a way to explicitly capture the voice of the customer. Using interactive visualization to address this big data enables you to do your analysis in an easily digestible, exploratory environment so that you can more easily interrogate, evaluate and understand what customers care about. Moreover, such high-performance visualization makes text analysis more accessible to people who don’t have training in analytics – they can simply point and click through the data to address whatever question strikes them.

Unstructured data analysis can be a daunting task. But with a single integrated, flexible analytic environment, you can readily visualize all your data – expanding the reach and insights of your analysis. High-performance visualization extends insights beyond traditional structured data. Interactive graphs, charts and other visuals can bring life to call center notes, trouble tickets, surveys, document stores and social data. Unstructured data will continue to grow exponentially. And those companies that use text analysis technology will have the advantage.

\textsuperscript{16} Walgreens, founded in 1901 in Chicago by Charles R. Walgreen, has grown to be the largest drug retailing chain in the US.
For More Information

Learn more about SAS Visual Analytics: sas.com/va

Discover how SAS Text Analytics can help your organization: sas.com/text-analytics
About SAS

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Through innovative solutions, SAS helps customers at more than 65,000 sites improve performance and deliver value by making better decisions faster. Since 1976 SAS has been giving customers around the world THE POWER TO KNOW®. For more information on SAS® Business Analytics software and services, visit sas.com.