Gaining full value from survey text

Make text analysis easier and more efficient with IBM SPSS Text Analytics for Surveys

Contents

Introduction

People communicate in many ways. The words they choose to use can tell you a lot about them and their experiences with your organization. For example, financial services firms regularly survey hundreds of thousands of customers to discover if certain banking products are meeting their needs. And universities often survey students to evaluate how they feel about particular courses and faculty, as well as to predict which students are likely to stay in school and complete their degrees. As data collection has evolved from paper surveys to online surveys, the sheer volume of text data has skyrocketed. While there's great value in using open-ended survey questions that let people express their views in their own words, your organization may find it tedious, time-consuming and costly to categorize or "code" text responses. The challenge is finding the most efficient way to extract that value.

IBM SPSS products enable organizations to analyze written communications, or text, more reliably and efficiently. We have also pioneered the development of solutions to support survey research about people's characteristics, attitudes, behaviors and beliefs. We have focused a great deal of effort on combining both types of information so that your organization can obtain a more complete view of the people you are surveying – whether they are customers, students or the general population. Your organization may then use this information to predict future needs or actions, and make smarter decisions.

This paper offers a brief review of the role of text in survey research, and then provides an overview of the steps you follow to conduct text analysis with IBM SPSS Text Analytics for Surveys. You can rely on this product to make text analysis easier and more efficient for your organization because it is a linguistics-based solution specifically designed for categorizing or "coding" text responses and sentiments.

Highlights:

This paper offers a brief review of the role to text in survey research, as well as an overview of approaches to text analysis.

It also describes the steps you take to conduct text analysis with IBM SPSS Text Analytics for Surveys, and how this software enables you to combine the analysis of text with the analysis of other survey data.

The role of text in survey research

Although text is difficult and time-consuming to analyze, text responses complement other data, providing more varied and detailed information about what respondents think, feel and do.

There are two fundamental reasons for including text responses. First, the words respondents choose often give you new insight into their thinking. Second, if you rely exclusively on closed-ended questions, you are framing not only the question but also the possible answers – in effect, constructing and interpreting reality for respondents. How can you be sure you've gotten it right? What if you have omitted a significant alternative response? Or what if the way a question has been asked has skewed or biased responses?

To acknowledge these issues, it's wise to include open-ended questions in your surveys. Previously, you may not have done more with these responses than select one or two to illustrate trends shown in the more easily quantified numeric data. Now, thanks to new tools for text analysis, you can more easily derive full value from text responses.

The most effective of these new tools use the technologies of linguistics-based text analytics. The differences between linguistics-based text analytics and other approaches are summarized in the next section of this paper. The most important difference is that linguistics-based text analytics is built upon a class of algorithms that analyze the structure and meaning of the language of a text – thus enabling computer systems to analyze the ambiguities inherent in verbal communication. Such linguistics-based text analytics technologies are the foundation upon which all IBM SPSS text analytics solutions are built, including IBM SPSS Text Analytics for Surveys.

Technologies relying on linguistics-based text analytics are most useful for allowing computer systems to analyze the ambiguities that are part of verbal communication.

Approaches to text analysis

There are several other approaches to text analytics. The challenge is to find the approach with the right balance between ease of use, reliability and efficiency.

Manual methods require you or your researchers to read a representative sample of text responses and create a set of appropriate categories for responses. A detailed list of coding instructions, or code frames, must be developed so that you can group responses consistently. Coding text responses manually can take days, even weeks, depending upon the number and complexity of text responses. Manual coding can be expensive and the process may also delay the delivery of needed information.

Approaches to text analysis include manual methods, automated solutions based on statistics and automated linguistics-based solutions.

A second approach is to employ automated solutions based on statistics. Some of these, however, simply count the number of times terms occur and calculate their proximity to related terms. Because they cannot factor in the ambiguities in human languages, relevant relationships may be hidden in masses of irrelevant findings – or missed altogether. Some of these statistics-based solutions compensate by providing ways for analysts to create rule books that help suppress irrelevant results. But these rulebooks need to be created and continually updated by analysts, which adds cost and complexity.

Other statistics-based solutions rely on self-learning tools such as Bayesian networks, neural networks, support vector machines (SVM) and/or latent semantic analysis (LSA). While these solutions can be more effective than other statistical approaches, they have the drawback of being "black boxes" – that is, using hidden mechanisms that cannot be adjusted except by highly skilled statisticians or programmers.

Automated linguistics-based solutions, by contrast, consider both grammatical structures and meaning when analyzing text. These solutions are based on the field of study known as natural language processing (NLP) or computational linguistics, a field that has been growing in importance as computing capabilities reached the level needed to analyze the ambiguities inherent in human language. Linguistics-based text analytics offers the speed and cost effectiveness of statistics-based systems but provides more reliable and useful results.

A key advantage of IBM SPSS Text Analytics for Surveys is that it gives you the ability to analyze respondents' attitudes and opinions. As a result, you gain a clearer understanding of what people like or don't like – and why. When you understand what people think and feel in their own words, you can draw more reliable conclusions about their future behavior and use that predictive insight to meet their needs more successfully.

Linguistics-based text analytics technologies are the foundation for IBM SPSS Text Analytics for Surveys. The features and interface of this product have been specifically designed for anyone who conducts survey research, whether your survey responses are short phrases or a few hundred words long. You can easily and efficiently import text responses, extract concepts or terms, group them by category and then export results either as categories or as dichotomies for analysis with other survey data.

In this way, textual data gains analytic value. Insight gained from text analysis can be used to complement other data analyses, enabling your organization to realize the benefits of predictive analytics. Whether you're a student, a business user, a market researcher or a decision maker for your organization, you can better anticipate future attitudes and behavior by uncovering patterns and trends in text – what we call predictive text analytics.

Preparing to use IBM SPSS Text Analytics for Surveys

To be successful in analyzing survey text responses, you must weigh many factors. These include:

- Survey text analysis, like any kind of text analytics, should be performed with clear objectives in mind. When planning a survey, determine what the goals of the study are and how text responses help in achieving these goals.
- The quality of the open-ended questions asked affects the usability of the responses received. Although Text Analytics for Surveys was designed to address a broad range of responses, avoiding questions that are very broad improves the relevance of responses and the resultant categories.
- Text analysis is not an exact science. There is no single "correct" outcome. Manual coding is subjective – it's influenced by your interpretation of the respondent's message. Two competent people can analyze the same data and reach different conclusions, depending on their individual perspectives. The linguistic technologies underlying Text Analytics for Surveys can, however, reduce the gap between individual interpretations.
- Text analysis is an iterative process. When you work with survey responses, you'll likely re-extract concepts and re-categorize responses using different category definitions or coding schemes, different term or synonym definitions, or different groupings of responses. You may repeat this process several times before you're satisfied with the results. Even so, the automation available with Text Analytics for Surveys will provide faster categorization – and the potential for more sophisticated analysis – than manual methods.

Steps in survey text analysis

With this latest version of Text Analytics for Surveys, you can quickly import survey data, extract key concepts, categorize responses and refine the results. Once you have categorized your data, you can export your categories for import into quantitative analytic tools, such as IBM® SPSS® Statistics, for further analysis and graphing. The following is a summary of the typical work flow that you will follow while using Text Analytics for Surveys.

Create projects easily with the Project Wizard

It's easier than ever to create a project because our new Project Wizard takes you step by step through the process, including the selection of Text Analysis Packages (TAPs). To begin:

- Import survey data, including open-ended responses, an ID variable and other reference variables. Data can be read from IBM SPSS Statistics data files, Microsoft® Excel®, any ODBC-compliant database program or an IBM® SPSS® Data Collection data source. Translate your non-English data into English data withsoftware available from third-party vendors.
- Select the categories and resources that you'd like to use in your TAP. Text Analytics for Surveys includes built-in TAPs (in English only) for customer, employee and product satisfaction surveys, so you save time and effort while gaining a key competitive advantage. You may also use customized TAPs from an earlier project. Or, if you don't have any pre-defined categories or specific dictionaries to use, you can create your own and then ask the software to refine them.

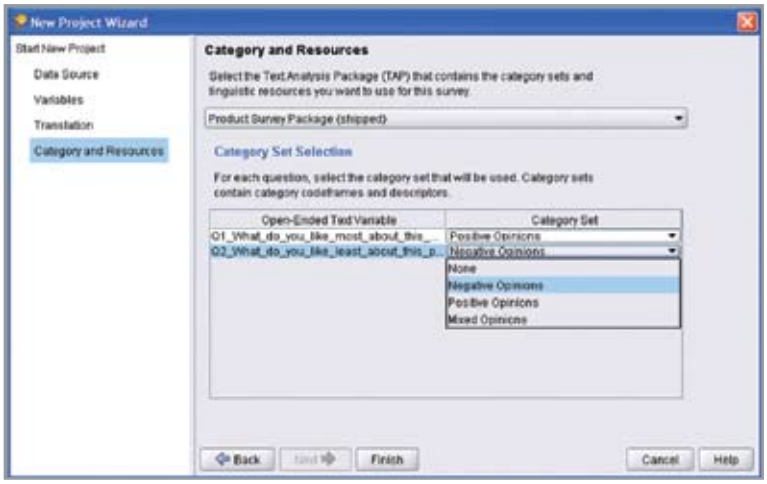

Figure 1: When creating a new project, the Project Wizard helps you select text fields and variables to be analyzed. It also shows you how to translate responses into English (if needed) and select certain TAPs. In this screenshot, the user has chosen the embedded Product Satisfaction TAP.

Create categories by:

- Importing an existing list of categories
- Dragging and dropping certain terms or named entities into each category
- Building them from scratch using the fully automated techniques in IBM SPSS Text Analytics for Surveys

Extract and index text

Although you may intervene in several ways during the extraction and classification processes, the extraction process in Text Analytics for Surveys doesn't require you to do so. The extraction process consists of six major steps:

1. Conversion of input data to a standard format
2. Identification of candidate terms (words or groups of words that identify concepts in the text)
3. Identification of equivalence classes (the base forms of candidate terms) and the integration of synonyms
4. Identification of named entities such as names of organizations, products, locations, etc.
5. Indexing
6. Sentiment analysis to detect positive or negative opinions

The libraries and dictionaries that constitute the linguistic resources of Text Analytics for Surveys have been especially designed for analyzing survey text. Some of these resources can be modified, and you can create custom libraries that improve the accuracy of your extraction. For a more detailed description of the linguistic technologies underlying Text Analytics for Surveys' extraction process, please see the white paper *Mastering New Challenges in Text Analytics*, which is available from your IBM SPSS sales representative.

Build categories

In text analysis, the term "categories" refers to a group of closely related concepts, opinions or attitudes. To be useful, a category should also be easily described by a short phrase or label that captures its essential meaning.

For example, if you are analyzing responses from consumers about a new laundry soap, you can create a category labeled scent that contains all of the responses describing the smell of the product. However, such a category would not differentiate between those who found the smell pleasing and those who found it offensive or too strong. Since Text Analytics for Surveys is capable of extracting opinions, you could then replace it by two other categories to identify respondents who enjoyed the scent and respondents who disliked the scent.

There are several ways to create categories. You can import an existing list of categories if you have one. For example, you may have a list of categories in an Excel spreadsheet, or maybe you've done manual categorization or used a third-party application previously and you'd like to reuse that list of categories.

Another option is to manually create categories. You can drag and drop certain terms or named entities into each category. Use this approach if you don't necessarily want to categorize all the responses you receive, but you really want to identify those that match a restricted set of categories. IBM SPSS Text Analytics for Surveys can help you by automatically extending your category definitions. For example, if you have selected apple as a descriptor of your category, the software can extend it to fruits or to apple pies according to your preferences.

If you're uncertain of what content you want to use for your categories and don't want to be biased, another option is building categories from scratch using the fully automated techniques in Text Analytics for Surveys. Its linguistics-based classification techniques are used to group noun terms and opinions. They create categories by identifying terms that are likely to have the same meaning or are either more specific or more general than the category represented by a term.

One advantage of using Text Analytics for Surveys is that it relies on a number of automated techniques for building categories – so you don't distort your results because you only used one or two techniques. The software's many techniques include term derviation, lexical series, semantic networks and co-occurence rules[1].

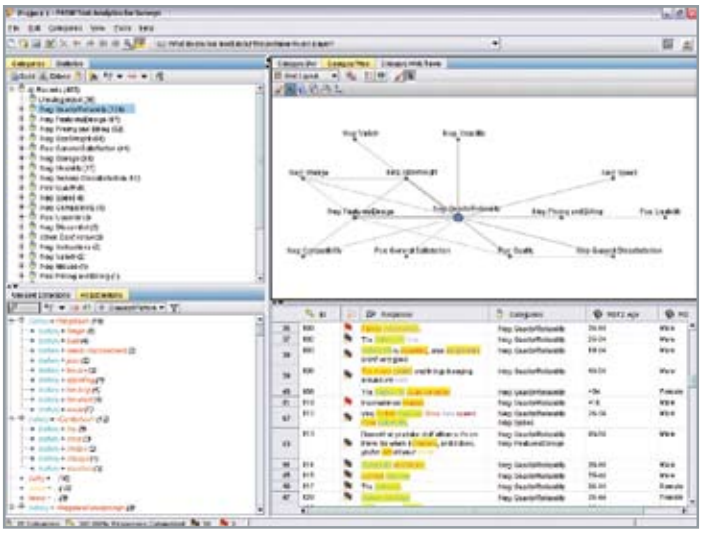

Figure 2: Responses have been automatically categorized using the Product Satisfaction TAP. All categories are listed in the left upper pane. Categorization has been performed based on results from concept extraction and sentiment analysis (left pane). Responses matching the category describing negative opinions about quality and reliability can be visualized very easily (lower right) and so can associated categories (upper right).

[1] For more information about concept grouping techniques, please use your Support login to access the IBM SPSS Text Analytics for Surveys 3.0 User's Guide, pages 9-10. If you don't have a Support login, another option is to download our technical report, *Mastering New Challenges in Text Analytics.*

Refine your resources for reliability

When IBM SPSS Text Analytics for Surveys uses the same data, with the same linguistic resources, it will always reproduce a prior analysis perfectly – it is 100 percent reliable.

This does not mean that there will be no errors in the analysis; however, you can now shift your focus from coding to fine-tuning. When humans code responses, they read them and capture all of the nuances of a statement (even if they have trouble applying the coding categories). IBM SPSS Text Analytics for Surveys can apply the coding categories, but the categories have to be defined so that nuances and distinctions are captured. The next steps in your analysis process include fine-tuning your linguistic resources, refining your category definitions and visualizing relationships between categories.

Fine-tune your linguistic resources

IBM SPSS Text Analytics for Surveys will automatically create categories for you. In order to capture all of the information in the responses, though, you will need to improve the program's linguistic base so that its category creation becomes more and more tuned to the idiosyncrasies of the text. To improve this base, you can customize and fine-tune the linguistic resources used in extracting from the text.

Fine-tuning, in this case, involves adding words to various linguistic libraries and dictionaries, specifying words to be excluded from the analysis, defining synonyms or creating custom libraries with a specific goal in mind. This goal is to accurately capture the ideas of the respondents in the text and remove ambiguity in the results.

Review your category definitions

In addition to refining the linguistic resources, you should review your categories by looking for ways to combine or clean up their definitions as well as checking some of the categorized responses. You can use the automated classification techniques to create your categories; however, you will surely want to perform a few tweaks to these definitions.

After using a technique, a number of new categories appear in the window. You can expand the categories to see the concepts that define each category. You can then review the responses in a category and make adjustments until you are comfortable with your category definitions.

No single automatic technique will perfectly code your response data. We recommend finding and applying one or more automatic techniques that work well with your data. After applying these techniques, review the resulting categories. You can then use manual techniques to make minor adjustments, remove any misclassifications or add records or concepts that may have been missed.

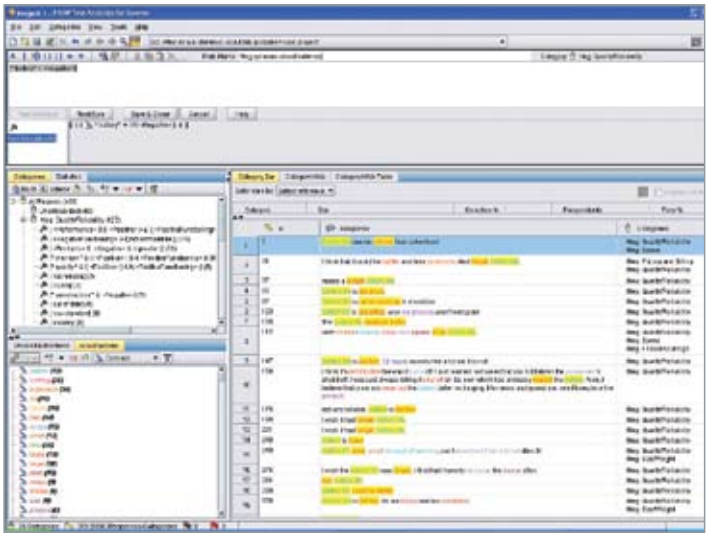

Figure 3: A new rule is manually added to improve the definition of the category called *negative opinions quality/reliability*. Records matching on negative opinions about batteries are displayed and will then be properly categorized.

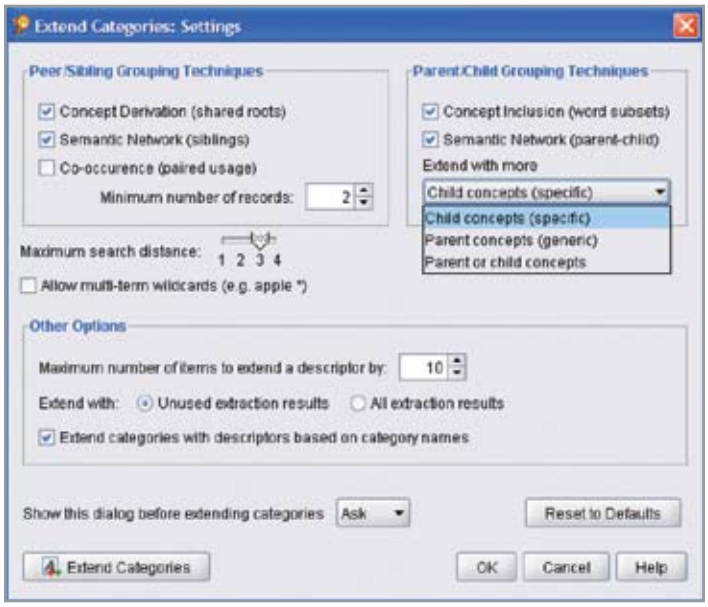

Figure 4: You can also improve category definitions by applying automated grouping techniques to better match text responses. Concept inclusion and semantic networks can be used, for example, to place apples, oranges and fruits into the same category.

Export your results for further analysis

Sometimes, the creation of categories of text responses is the only analysis that a particular survey requires. Knowing the major themes expressed by respondents, and how many respondents mentioned each theme, may be enough to provide insight into respondents' attitudes, behaviors or beliefs.

At other times, though, you may want to perform additional reporting and analysis. It may be beneficial, for example, to create tables and graphs displaying survey results. You may want to use variables from other sections of the survey questionnaire to shed further light on text respondents, or analyze the categories found in text responses along with other survey data. IBM SPSS Text Analytics for Surveys enables you to carry out additional analyses by exporting text categories as dichotomies either to Statistics or to Excel. In either of these programs, you or your organization's researchers can perform statistical calculations and create graphs showing relationships contained in the data.

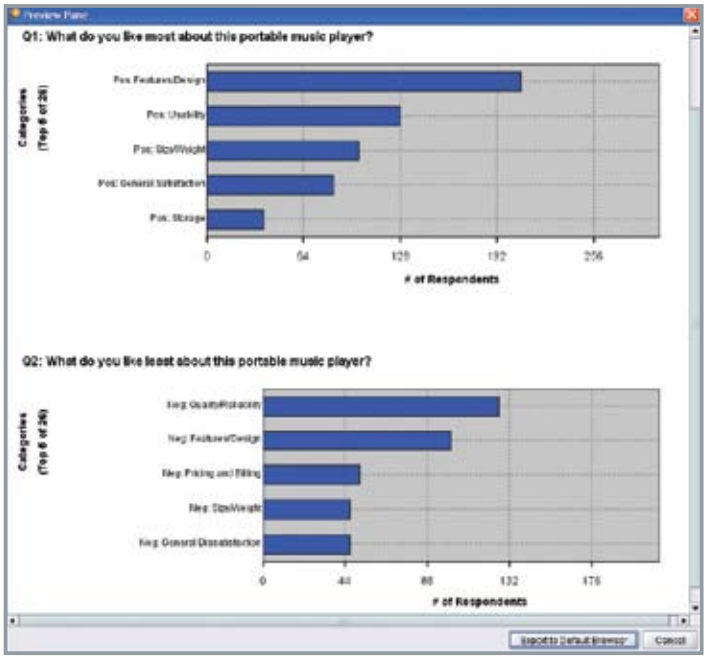

Figure 5: Results generated by Text Analytics for Surveys can be easily exported as pictures or HTML files – so you can include them in presentations.

Categories or codes created with Text Analytics for Surveys can be saved for reuse in similar or follow-up studies. This program can also exchange data through the IBM® SPSS® Data Collection Data Model used by our feature-rich IBM SPSS Data Collection family of survey research products. This family includes products that support the creation and deployment of sophisticated paper, telephone and online questionnaires, as well as their translation into multiple languages. IBM SPSS Data Collection products enable users to perform advanced data analyses and share or publish results efficiently and cost effectively.

Conclusion

This paper provided a brief review of the role of text in survey research as well an overview of approaches to text analysis. It then described the steps you follow to conduct text analysis with IBM SPSS Text Analytics for Surveys. Whether you're a student in the social sciences or statistics, a business analyst or a market researcher, this program makes it easier for you to:

- Recode open-ended text responses
- Translate responses from other languages into English, using third-party tools
- Reuse previous projects to automatically recode new surveys
- Produce consistent results and reports

At the same time, you'll spend less time on manual work and reap increased cost savings.

Because the techniques available with Text Analytics for Surveys enable you to combine the analysis of text with the analysis of other survey data, your organization gains a richer, more detailed understanding of your results than is possible with other methods.

By making text responses more easily quantifiable, Text Analytics for Surveys opens the door for incorporating insight gained from text into quantitative analyses, including the kind of predictive analysis that is possible with IBM SPSS data mining and decision optimization solutions. This means that Text Analytics for Surveys can be a key component of other research, enabling your organization, for example, to use survey research data to deepen its understanding of customers, employees or constituents; anticipate changing needs; and prepare to meet them successfully.

About IBM Business Analytics

IBM Business Analytics software delivers complete, consistent and accurate information that decision-makers trust to improve business performance. A comprehensive portfolio of business intelligence, predictive analytics, financial performance and strategy management, and analytic applications provides clear, immediate and actionable insights into current performance and the ability to predict future outcomes. Combined with rich industry solutions, proven practices and professional services, organizations of every size can drive the highest productivity, confidently automate decisions and deliver better results.

As part of this portfolio, IBM SPSS Predictive Analytics software helps organizations predict future events and proactively act upon that insight to drive better business outcomes. Commercial, government and academic customers worldwide rely on IBM SPSS technology as a competitive advantage in attracting, retaining and growing customers, while reducing fraud and mitigating risk. By incorporating IBM SPSS software into their daily operations, organizations become predictive enterprises – able to direct and automate decisions to meet business goals and achieve measurable competitive advantage. For further information or to reach a representative visit www.ibm.com/spss.

IBM Corporation
Route 100
Somers, NY 10589

Produced in the United States of America
May 2010

Please Recycle

Business Analytics software

YTW03030USEN-01