Analyzing survey text: a brief overview

Learn how IBM SPSS Text Analytics for Surveys gives you greater insight

Contents

Introduction

Although people communicate in many ways, perhaps the most powerful and pervasive way is through the words they choose. Whether verbal communications take the form of carefully crafted written reports, text responses to survey questions, or informal, even cryptic, notes or emails, they can offer great informational value. The challenge lies in finding a way to extract that value.

SPSS was a pioneer in developing tools 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. For this reason, we have focused a great deal of effort not only on the statistical tools that analyze responses to closed-ended survey questions but also on techniques to extract and categorize the wealth of information contained in verbatim responses to open-ended questions. By combining both types of information, organizations obtain a more complete view of the people they are surveying–whether these are customers, students, or the general population.

As a new member of the IBM organization, SPSS brings its leading-edge analytical tools, making IBM SPSS technology the global leader in predictive analytics. IBM SPSS offerings include industry-leading products for data collection, statistics, data and text mining, and text analysis to allow your organization to analyze the responses of open-ended surveys.

IBM SPSS tools are based on industry standards and can easily integrate with your existing infrastructure to improve accuracy, decrease manpower and minimize loss. The combined effort of IBM and SPSS brings you the utmost in flexibility in the kinds of data you mine and how you deploy results.

Organizations can then use this information to anticipate or predict future needs or actions. This paper provides a brief review of the role of text in survey research. Then it describes several approaches to analyzing survey text and provides a look at the workings of IBM SPSS Text Analytics for Surveys, a linguistics-based solution specifically designed for categorizing or "coding" survey text responses.

The role of text in survey research

If text is so difficult and time-consuming to analyze, why not just avoid using questions that require text responses? The answer, as experienced researchers know, is that 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 researchers new insight into their thinking. Second, if researchers rely exclusively on closed-ended questions, they are framing not only the question but also the possible answers–in effect, constructing and interpreting reality for respondents. How can researchers be sure they've gotten it right? What if they have omitted a significant alternative response? Or what if the way a question has been asked has skewed or biased responses? In acknowledgement of these reasons, many researchers incorporate open-ended questions in their surveys. But they may not do 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, researchers can more easily derive full value from text or verbatim responses. The most effective of these new tools use the technologies of linguistics-based text mining. The differences between linguistics-based text mining and other approaches are summarized in the next section of this paper. The most important difference is that linguistics-based text mining 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 mining technologies are the foundation upon which all IBM SPSS text mining solutions are built, including IBM SPSS Text Analytics for Surveys.

Approaches to text mining

There are several other approaches to text mining. The challenge is to find the approach with the right balance between reliability and efficiency. Manual methods require that researchers or coding supervisors read a representative sample of text responses and create a set of categories into which coders can place responses appropriately. A detailed list of coding instructions, or code frames, must be developed so that individuals doing the coding can group responses consistently. Although human coders usually find it easy to understand the content of a response, they may disagree on how it should be categorized. This divergence reduces the reliability of the research results.

Even with a minimum amount of disagreement, coding text responses manually can take days, even weeks, depending upon the number and complexity of text responses. This is not only costly but, in some cases,

delays the delivery of needed information. There are several automated methods for analyzing text that are designed to speed up the process and minimize the amount of human intervention required.

One approach is to employ solutions based on statistical formulas, neural networks, and other techniques, often used in combination with detailed rulebooks. Considerable expertise is required to develop and use such methods, which is one reason many organizations do not implement them. Another reason is that these solutions seem like a "black box," with their underlying technologies hidden from users and beyond their control. Another problem with statistics-based automated solutions is that while they can process text quickly, their accuracy is fairly low. Many statistics-based solutions simply count the number of times terms occur and calculate their proximity to related terms.

If survey text responses are fairly simple–lists of preferred cereal brands or pizza toppings, for instance–a statistics-based solution may provide useful results. With even slightly more complex text responses, however, statistics-based solutions produce many irrelevant results and miss results they should have found. Statistics-based solutions treat text as a "bag of words," rather than incorporating the structure and meaning of the language in their analysis.

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 mining offers the speed and cost effectiveness of statistics-based systems but provides more reliable and useful results.

Advanced linguistics-based text mining technologies provide the foundation for IBM SPSS Text Analytics for Surveys. However, the features and interface of IBM SPSS Text Analytics for Surveys have been specifically designed with survey researchers in mind. Survey researchers can easily and efficiently import text responses, extract concepts or terms, group them by type, and then export results either as text or as dichotomies for analysis with other survey data.

In this way, textual data gains predictive value. Researchers and decision-makers can better anticipate future attitudes and behavior by uncovering patterns and trends in text–what we like to call IBM SPSS predictive text analytics. In addition, insight gained from text analysis can be used to complement other data analyses, enabling organizations to realize the benefits of predictive analytics.

Steps in survey text analysis

Survey responses are normally relatively short–a phrase, sentence, or short paragraph. IBM SPSS Text Analytics for Surveys was designed for this type of text, but it can also process responses that are several hundred words in length. The major steps involved in the analysis process are extraction and categorization. During both steps, IBM SPSS Text Analytics for Surveys allows researchers to modify the underlying technology to best meet the needs of their particular survey project. Before describing how extraction and categorization work in IBM SPSS Text Analytics for Surveys, a few words about the process as a whole are in order.

Preparing to use IBM SPSS Text Analytics for Surveys

To be successful, a researcher preparing to analyze survey text responses must weigh many factors. These include the following:

- Survey text analysis, like any kind of text mining, should be performed with clear objectives in mind. When planning a survey, a researcher should 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. 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. Text mining is subjective–influenced by the analyst's interpretation of the message conveyed by the respondent. Two competent people can analyze the same data and reach different conclusions, depending on their individual perspectives. The linguistic technologies underlying IBM SPSS Text Analytics for Surveys can, however, reduce the gap between individual interpretations.
- Text analysis is an iterative process. In working with survey responses, a research analyst is likely to re-extract concepts and re-categorize responses using different category definitions or coding schemes, different term or synonym definitions, or different groupings of responses. This process may be repeated several times before the results are satisfactory. Even so, the automation available with IBM SPSS Text Analytics for Surveys will provide faster categorization–and the potential for more sophisticated analysis–than manual methods.

What researchers see when using IBM SPSS Text Analytics for Surveys

Some aspects of the extraction and classification processes in IBM SPSS Text Analytics for Surveys occur automatically and without the need for user intervention. The screenshots on this page provide a view of what a researcher might see while coding a set of text responses.

Survey responses imported.
Researchers can import text responses from a variety of sources, including Microsoft® Excel™; IBM® SPSS® Statistics; IBM® SPSS® Data Collection Data Model, which underlies IBM SPSS survey research products such as IBM® SPSS® Data Collection Interviewer and any ODBC-compliant database.

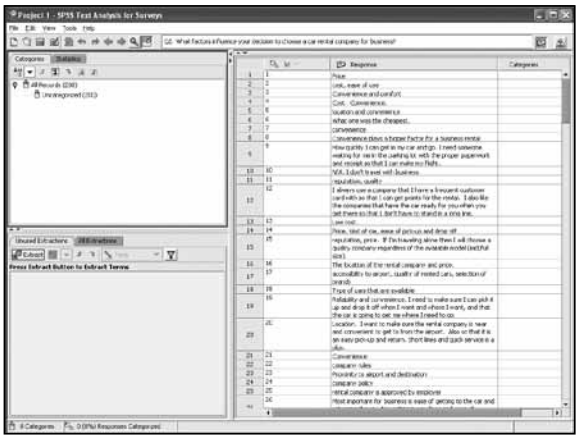

Figure 1: Responses to an open-ended question from a survey studying factors that influence consumers' car rental decisions are shown here.

Terms and concepts extracted.
After a user clicks the Extract button, IBM SPSS Text Analytics for Surveys automatically extracts the relevant terms and concepts contained in responses to one question. Manual intervention is not required. Users of IBM SPSS Text Analytics for Surveys can, however, modify some of the product's dictionaries and libraries so that extraction is carried out with even greater precision.

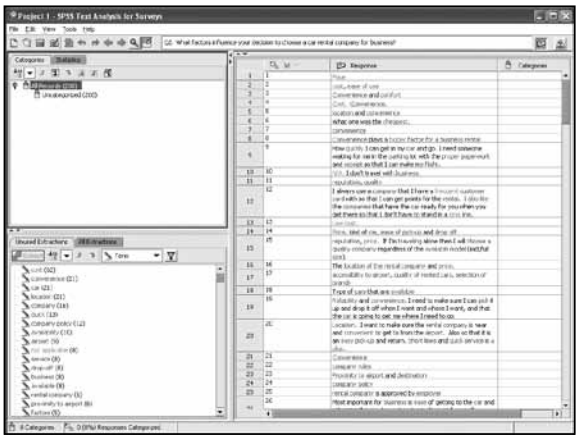

Figure 2: Onscreen, a researcher or coder sees the full text of responses in the right-hand pane, with extracted terms highlighted in color and listed in the pane at lower left.

Responses classified and categorized.
Through an easy-to-use dialog box, a researcher can choose from a variety of techniques when categorizing extracted terms.

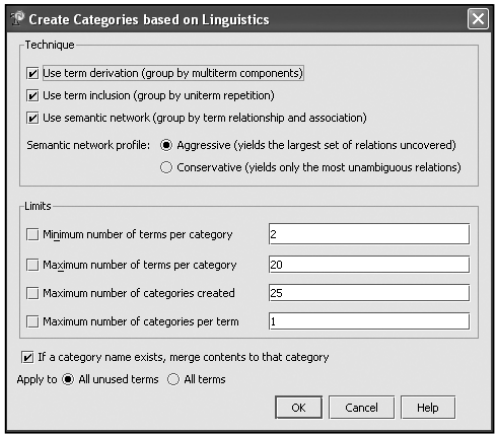

Figure 3: Term derivation, term inclusion, and semantic network are three linguistics-based techniques that users of IBM SPSS Text Analytics for Surveys may select, separately or in combination, when creating categories. The default setting uses all three together.

Categories manually modified.
Users see not only the response text and the extracted terms and categories but also a graphic visualization of the relationships between responses.

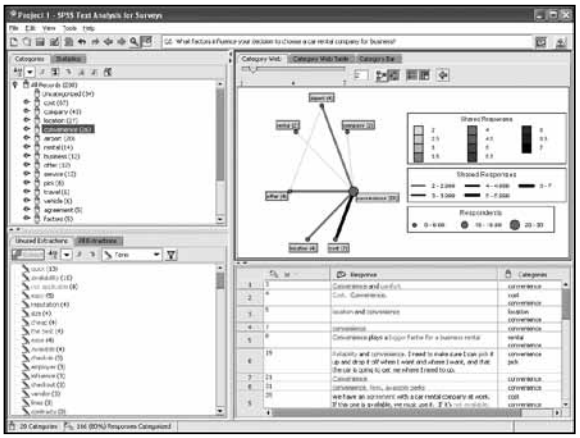

Figure 4: The ability to see a map of the co-occurrence of responses helps researchers or coders fine-tune categories, resulting in more precise groupings of responses. Categories can be saved and reused when coding similar or follow-up surveys.

Results exported for analysis

Once users are satisfied with the categorization of responses, they can export results either as dichotomies or as text categories. These can be used to create graphs and tables describing the analysis of text responses, either separately or in combination with other survey data.

Figure 5: Results can be exported to IBM SPSS Statistics or Excel for analysis or to create a variety of graphs that help communicate survey findings.

How extraction works

Although users can intervene in several ways during the extraction and classification processes, the extraction process in IBM SPSS Text Analytics for Surveys does not require user intervention. 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. Type assignment
5. Indexing
6. Link analysis, and event and sentiment extraction

The libraries and dictionaries that constitute the linguistic resources of IBM SPSS Text Analytics for Surveys have been optimized for analyzing survey text. Some of these resources can be modified, and researchers can create custom libraries suited to their particular topic or type of research. This customization results in a more useful initial list of extracted terms and concepts, so that fewer iterations and less manual intervention are required during the categorization process. Once created, custom libraries can be shared with others.

For a more detailed description of the linguistic technologies underlying IBM SPSS Text Analytics for Surveys' extraction process, please visit www.ibm.com/software/analytics/

As terms are extracted, they are assigned to a type. A type is a semantic grouping of terms. Types include such things as higher-level concepts; positive and negative terms and qualifiers; qualifiers that depend upon context; and names of people, places, or organizations.

IBM SPSS Text Analytics for Surveys also embeds a list of patterns as part of the extraction process. Patterns are a combination of terms and types. They are particularly useful when attempting to discover opinions about a particular topic, such as what negative statements respondents may have made. One of the strengths of IBM SPSS Text Analytics for Surveys is that it enables researchers to easily distinguish between positive and negative responses: "Thought the class was great" versus "Thought the class was awful," for example.

How classification and categorization work

The next step is the classification and categorization of text responses. *Classification* refers to the generation of category definitions through the use of built-in techniques. Categorization refers to the scoring or labeling process during which unique identifiers are assigned to each response.

Because the product has robust classification algorithms, researchers using IBM SPSS Text Analytics for Surveys can automatically create categories or codes into which survey responses are placed. Although researchers could use the terms, types, and patterns uncovered in the extraction process to create categories manually, the product's automated methods do this far more quickly and easily. Both linguistics-based and statistics-based classification techniques are available.

Researchers can also use a combination of automated and manual techniques to finalize categories. Because every dataset is unique, the choice of techniques and the order in which a researcher applies them are likely to vary from one project to another. In all cases, however, the process is iterative: a researcher applies certain techniques, evaluates the results, makes changes either to the technique chosen or to the resultant categories, and further refines the categories.

Here is a brief description of the automated classification techniques available with IBM SPSS Text Analytics for Surveys.

Linguistics-based techniques

IBM SPSS Text Analytics for Surveys' linguistics-based classification techniques are used to group noun terms. They create categories by identifying terms that are likely to have the same meaning (also called *synonyms*) or are either more specific than the category represented by a term (also called *hyponyms*) or more general (*hyperonyms*). For cleaner results, these linguistic techniques exclude adjective terms and other qualifiers.

"IBM SPSS Text Analytics for Surveys is exactly what the university has been searching for to extract critical nuggets of information from long text survey responses. Using this new product, we were able to quickly identify concerns students had about university services. These concerns were not apparent in the quantitative survey data alone. However, when we put IBM SPSS Text Analytics for Surveys to work analyzing the open-end text responses, we were able to uncover those concerns, and act on them. The enhanced insights that IBM SPSS Text Analytics for Surveys provides will significantly improve the quality of our survey analysis and the decision-making ability of our organization."

– John Lemon
Senior Computing Advisor
University of Aberdeen, Scotland

Term derivation. This linguistics-based classification technique creates categories by grouping multiple-word terms whose components have related word endings (also called *suffixes*). This technique is very useful for identifying synonymous multiple-word terms, since the terms in each category generated are synonyms or closely related in meaning. Term derivation works with response data of varying lengths and generates a small number of compact categories.

Lexical series. This technique, based on term inclusion, creates categories by taking terms and finding other terms that include them. A lexical series based on term inclusion often corresponds to a taxonomic hierarchy (a semantic "is a" relationship). For example, the term *sports car* would be included in the term *car*. In IBM SPSS Text Analytics for Surveys, one-word or multiple-word terms that are included in other multiple-word terms are examined first and then grouped into appropriate categories. Term inclusion works with response data of varying lengths and typically generates a larger number of categories than term derivation.

Semantic network. This technique creates categories based on an extensive index of word relationships. First, extracted terms that are synonyms, hyponyms, or hyperonyms are identified and grouped. IBM SPSS Text Analytics for Surveys uses a semantic network with sophisticated algorithms to filter out nonsensical results. This technique produces very good results when the terms are known to the semantic network and are not too ambiguous. It is less helpful when text contains a large amount of specialized, domain-specific terminology that the network does not recognize.

Statistics-based techniques

The primary statistical technique offered in IBM SPSS Text Analytics for Surveys is based on the frequency with which terms, types, or patterns occur. This technique can be used both on noun terms and or other qualifiers. *Frequency* refers to the number of records containing a term or type and all its declared synonyms.

Grouping items based on how frequently they occur can yield interesting results, since it may indicate a common or significant response. This approach tends to work best when the text data contain straightforward lists or simple terms. It can also be useful to apply this technique to any terms that are still uncategorized after other techniques have been applied.

Reliability and fine-tuning

When coding data, researchers want to be confident that the categories created accurately reflect respondents' answers. This means that two independent coders, using the same rules or code frame, will code the same responses identically. When text analysis is done manually, this is a critical issue. A valuable set of categories can be created, but if they cannot be reliably applied to the responses, their value decreases substantially. When IBM SPSS Text Analytics for Surveys is applied to the same data, using 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 initial categorization, but rather that the focus can now shift to something else–fine-tuning. In manual coding, the coders read the response and can capture all 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.

There are two ways that fine-tuning can be performed: by refining the linguistic resources and by refining the category definitions. IBM SPSS Text Analytics for Surveys easily creates categories without manual intervention, but these categories may not capture all the information in the responses. It is possible, however, to modify the program's basic linguistic resources so that category creation becomes increasingly sensitive to the idiosyncrasies of the text. This can be done by adding words and phrases to the software's linguistic libraries and dictionaries, by specifying words to be excluded from the analysis, by defining synonyms, or by creating custom libraries.

In addition to refining the linguistic resources, researchers can review categories to ensure that the ones created are relevant and that no important category has been omitted. If necessary, manual techniques can be used to make minor adjustments, remove any misclassifications, and add records or terms that may have been missed.

Analyzing text with subtlety and speed

IBM SPSS Text Analytics for Surveys is powered by advanced linguistics-based text mining technologies. These advanced, NLP-based technologies process text with a combination of subtlety and speed, so researchers can categorize text responses far faster than with manual methods. Researchers can modify dictionaries and libraries for even faster processing. Yet they still can intervene manually to fine-tune categories and the responses they contain.

Additional 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, behavior, or beliefs.

At other times, though, a researcher may want to perform additional reporting and analysis. It may be beneficial, for example, to create tables and graphs displaying survey results. A researcher 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 researchers to carry out additional analyses by exporting text categories as dichotomies either to IBM SPSS Statistics or to Microsoft Excel. In either of these programs, researchers can perform statistical calculations and create graphs showing relationships contained in the data.

Categories or codes created with IBM SPSS Text Analytics for Surveys can be saved for reuse in similar or follow-up studies. IBM SPSS Text Analytics for Surveys can also exchange data through IBM SPSS Data Collection Data Model. 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 Data Model enables users to perform advanced data analyses and share or publish results efficiently and cost effectively.

Conclusion

This paper has provided a brief review of the role of text in survey research. It then described several approaches to categorizing survey text and, in particular, the extraction and categorization processes of IBM SPSS Text Analytics for Surveys. Because the techniques available with IBM SPSS Text Analytics for Surveys enable researchers to combine the analysis of text with the analysis of other survey data, researchers can gain a richer, more detailed understanding of their results than is possible with other methods.

By making text responses more easily quantifiable, IBM SPSS 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 IBM SPSS Text Analytics for Surveys, a key component of IBM SPSS' predictive solutions family, can be a key component of other research or business organizations' approach to using survey research data to deepen their understanding of customers, anticipate changing needs, and prepare their organizations 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

YTW03100-USEN-02