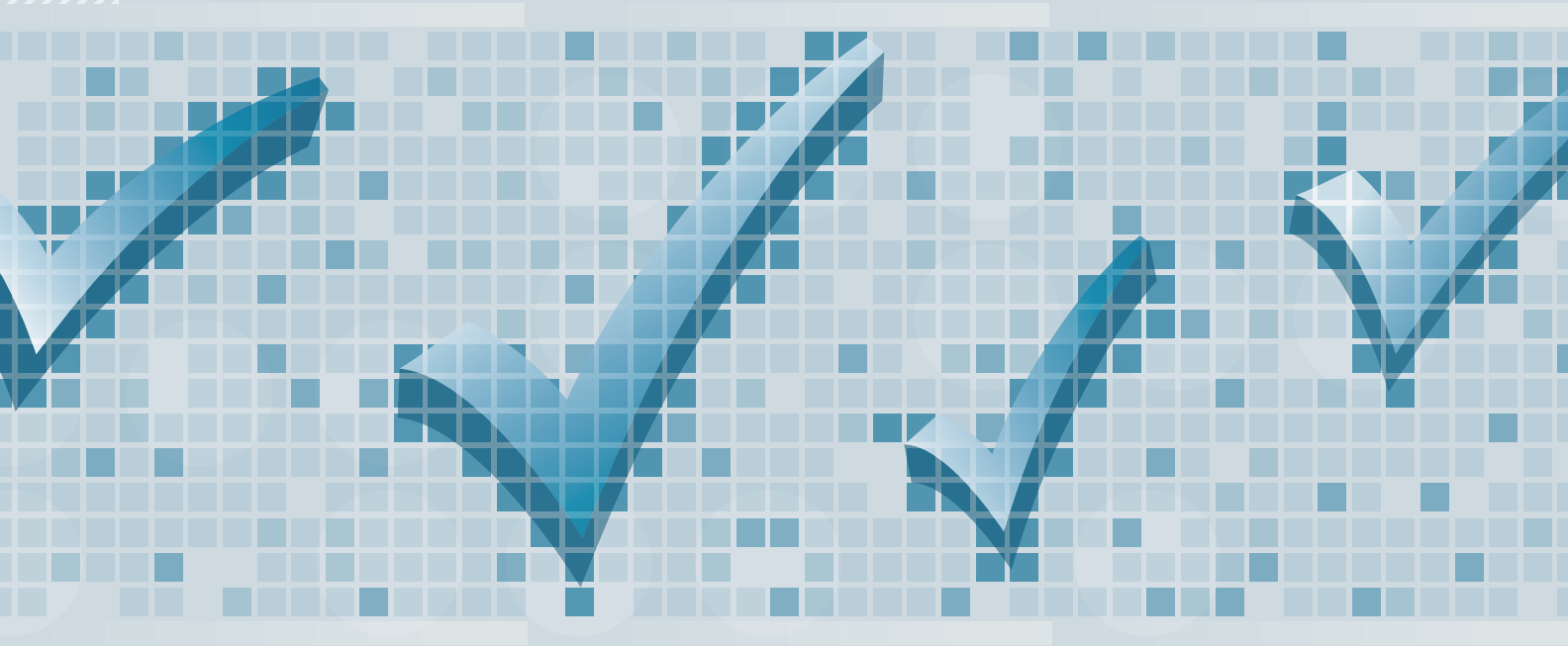
TDWI RESEARCH

TDWI CHECKLIST REPORT

HOW TO GAIN INSIGHT FROM TEXT

By Fern Halper



Sponsored by

Angoss
Predict. Act. Perform.™

LEXALYTICS
Read Between The Lines

sas

tdwi.org

tdwi

SEPTEMBER 2013

TDWI CHECKLIST REPORT

HOW TO GAIN INSIGHT FROM TEXT

By Fern Halper

TABLE OF CONTENTS

555 S Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org

FOREWORD

Text is everywhere—and its volumes are growing rapidly. Text comes from internal sources such as e-mail messages, log files, call center notes, claims forms, and survey comments as well as external sources such as tweets, blogs, and news. Text analytics, a technology used to analyze the content of this text, is rapidly gaining momentum in organizations that want to gain insight into their unstructured text and use it for competitive advantage. Factors fueling growth include a better understanding of the technology's value, a maturing of the technology, the high visibility of big data solutions, and available computing power to help analyze large amounts of data.

Text analytics is being used across industries in numerous ways, including customer-focused solutions such as voice of the customer, churn analysis, and fraud detection. In fact, many early adopters have used the technology to better understand customer experience, and this is still one of the most popular use cases. However, text analytics is also being used in other areas such as risk analysis, warranty analysis, and medical research.

The technology provides valuable insight because it can help answer questions involving *why* and *what*. For example, why did my customer leave? What is causing the increase in service calls? What are the best predictors of a certain risk? Text analytics can aid in discovery and improving the lift or accuracy of analytical models. This impacts an enterprise's top *and* bottom lines.

Companies are realizing that text is an important source of data that can improve and provide new insight. The questions these companies face include where to start and how to think about text as data. This Checklist Report focuses on helping organizations understand how to get started with text analytics, including:

- Basic definitions of text analytics
- How such analysis can add structure to unstructured data
- How unstructured data can be used and its importance to data discovery and advanced analytics
- How to think through a text analytics problem
- What resources are needed to get the maximum value from text

NUMBER ONE

IDENTIFY A PROBLEM WORTH SOLVING.

The first step in any analysis is to identify the problem you're trying to solve. This is also true for text analytics. It is important to start with the end goal in mind. It generally makes sense to pick an initial problem that has relatively high visibility and where it is fairly easy to get at the data. If possible, it should be a quick win that uses a proof of concept (POC). This accomplishes three objectives. First, it costs less and carries a lower risk than going all out. Second, a problem worth solving will earn a seat at the executive table, which can help to keep momentum high. Finally, a technical benefit of the POC is to ensure that the technology you're using works with your specific data.

An important facet of identifying the right problem is the ability to justify spending money on it as part of the business case. That means looking for a problem where improvements can be measured in one or both of the following ways:

- **Bottom-line improvements.** This generally means you will focus on reducing costs. For text analytics, this often involves decreasing the time required for a particular task (a productivity improvement). Bottom-line impacts can include reducing human misclassification errors of documents, which means finding information more quickly. Another example would be reducing costs by not having to read hundreds of thousands of survey responses and code them manually. Such potential improvements can be used in a text analytics business case.
- **Top-line improvements.** Top-line impacts are wide and varied. They might include improved customer retention rates or new customer win rates. They may also involve positive referrals or a decrease in defect rates or incidents of fraud.

Another factor to consider as part of the business case is whether the solution is multi-purpose. For example, there are numerous products on the market that use text analytics to gain insight into social media to understand customer opinions and sentiment. It is important to think beyond the first use case and consider your options wisely: i.e., point solutions versus more robust, integrated solutions.

NUMBER TWO

DETERMINE DATA REQUIREMENTS.

There are several issues to consider in terms of gaining access to the data as well as preparing it for analysis.

- **Data sources.** Data sources can be internal or external to your organization. For example, if a company is trying to predict churn, it may determine that internal call center notes are the most important source of text data. These call center notes might be stored somewhere inside the firewall or outside it (such as in the cloud). Part of identifying the data sources includes determining what spoken languages you need to include in the analysis.
- **Data access.** This includes gaining the right to use certain internal or inter-company data stores, which can be a hurdle, as well as being able to physically connect to the data. Will you access the data in real time via an API or get a one-time export to some common format such as comma-separated variable (CSV)? Accessing external data might require crawlers (programs that find and gather Web pages) or working with a data aggregator. What are the terms of service for the sites where you are gathering data? Typically, companies tend to deal with their internal data first, except when they use a specific social media analysis solution.
- **Data security.** Text data that contains personally identifiable or other highly sensitive information must be dealt with differently than a stream of public tweets.
- **Data timeliness.** Analyzing numbers once a quarter will require a different approach than analyzing data daily, hourly, or in real time.
- **Data preparation.** The most important aspects of this step are data normalization and data cleansing. Data normalization makes sure that the data acquired from each source will be able to match with other sources. Data cleansing deals with issues such as typos to ensure completeness of input (e.g., for social data) and that the data is trustworthy. Determining data quality for unstructured data is a science that is still evolving and can be time-consuming.

NUMBER THREE

IDENTIFY WHAT NEEDS TO BE EXTRACTED.

Text analytics is the process of analyzing unstructured text, extracting relevant information, and transforming it into structured information that can be leveraged in various ways. Text analytics can use a combination of natural language processing, statistical, and machine learning techniques. Entities, themes, concepts, and sentiment are all examples of structured data that can be extracted from text analysis. These are often referred to as "text features":

- **Entities:** often called *named entities*. Examples include names of persons, companies, products, geographical locations, dates, and times. Entities are generally about who, what, and where.
- **Themes:** important phrases or groups of co-occurring concepts about what is being discussed. A theme might be "women's rights" or "cloud computing." A particular piece of content might contain many themes.
- **Concepts:** sets of words and phrases that indicate a particular idea or meaning with which the user is concerned. A concept might be "business" or "smartphones." A particular piece of content generally is only "about" a few concepts.
- **Sentiment:** Sentiment reflects the tonality or point of view of the text. The concept "unhappy customer" would lead to a negative sentiment.

The goal is to accurately extract the entities, concepts, themes, and sentiment in which you are interested. Solutions offer various features out of the box. A vendor might include only a dictionary, list of names, or synonym list. Another might support hierarchical taxonomies to better organize information. The disadvantage of any purely list-based or taxonomic solution is that you're limited to finding what's in the list.

To address this issue, some vendors now incorporate statistical models based on machine learning into their solutions to help users extract features that were not preconfigured. Vendors that provide models often pre-train them so users don't need to do anything but simply use the model. Some vendors provide hybrid approaches (statistical and rules based), which provide the benefits of collection investigation combined with the specificity that comes from linguistic rules.

NUMBER FOUR

EXPLORE, DISCOVER, AND VISUALIZE.

One of the beauties of text analytics is that users can *discover* insights they didn't know about before. Visualization is important for both exploration and discovery in text analytics.

- **Visualization in text exploration.** It is important to explore text data to understand what is in it as well as to iterate on building out the text model (i.e., the relevant entities, themes, and concepts in Number Three). Exploration is a common first best practice even in structured data analysis. There are many ways to do this with text. For example, a term cloud is a simple way to help identify which words appear most often in the corpus; in this visualization type, the size of the word (or phrase) indicates its importance. Word frequency counts are another kind of visualization. Concept networks or links (sometimes called social networks or term networks) can also be useful. These links show the relationships between terms and entities. The thickness of the line determines the strength of the relationship. Exploring the data in this way is often an iterative and interactive process. (See Figure 1.)
- **Visualization for discovery.** Once the text model is developed and the entities, themes, concepts, and sentiments are extracted, visualization can be very powerful for discovery. In fact, the process of examining and interacting with text data alone (even without combining it with traditional structured data) can help end users make valuable discoveries. For example, using a vendor solution with strong visualization capabilities, text data can be useful for isolating issues or uncovering what customers are complaining about. A key feature to look for in this kind of visualization is the ability to easily interact with the data. Bar charts, tables, line charts, histograms, scatterplots, and other visualizations are important for this. Other features to look for include filtering the visualization to view an important characteristic, filtering by a time interval, or viewing specific concepts for a topic. In addition, the ability to drill down into detailed text behind a specific visualization is useful. (See Figure 2.)

Figure 1

Click to enlarge

Figure 2

Click to enlarge

NUMBER FIVE

CONSIDER SENTIMENT AND OTHER MEASURES.

Understanding people's opinions and sentiment is important for a broad range of use cases. Marketers want to understand how consumers feel about their brand or how they compare brands. Politicians want to understand how voters view them. Sentiment analysis can be tricky, especially when it comes to understanding irony, sarcasm, and other nuances of language. In addition, the meaning of a word or phrase is often context sensitive. For example, the phrase "This phone is sick" has a different sentiment from "Product XYZ made my daughter sick." Expect sentiment accuracy (i.e., correct classification) of between 60% and 80% depending on the quality of the content, and provided the vendor solution is robust.

Vendors generally use a scoring scheme to analyze sentiment. The granularity of the sentiment scoring and the actual scoring scheme can affect the sentiment score. Sentiment analysis can be done at multiple levels, including at the document, paragraph, sentence, and even the feature level. If you need to understand the sentiment directed at a brand, it is insufficient to simply get the sentiment of a document or sentence.

A key feature for sentiment analysis is the ability for the end user to be able to tune the underlying sentiment to match the use case. Ideally, the system can learn from any changes.

Because sentiment can be considered a polarity or an intensity measure, it also makes sense to think about metrics beyond sentiment. For example, it might be relevant to consider metrics such as mood or sentiment shift. Some vendors have recently begun to include other forms of polarity measures. For example, an HR-related text analysis might be interested in levels of expertise. A legal analysis might want to incorporate levels of truthfulness. These polarity measures will continue to evolve, so it is important to consider how flexible sentiment scoring is when assessing your requirements.

NUMBER SIX

THINK ABOUT ADVANCED ANALYTICS AND LIFT.

In addition to using text data in discovery, it can also be utilized in more advanced analytics.

- **Marrying structured and unstructured data.** Increasingly, companies are marrying text data with structured data in predictive analytics to potentially increase the lift (i.e., the effectiveness and accuracy) of a model. This is a popular use case for combining structured and unstructured data. For example, in churn analysis, companies combine the text from call center notes (which includes insight about why a customer called as well as sentiment associated with the call) with structured data such as demographic data. The concepts, entities, sentiments, and themes provide additional sources of attributes for a predictive model. Companies have found that this can help to improve the model's lift.
- **Predictive analytics using text alone.** Unstructured text data can also be utilized without structured data in predictive models. This is a newer kind of analysis that can provide significant insight. For example, entities, concepts, and themes can be clustered using statistical techniques for customer segmentation. In addition, companies can use survey comments to assign entities, concepts, and themes as data and use this for prediction without structured data.
- **Big data analytics.** Text data is also an important component of big data analytics. For instance, companies across numerous industries are utilizing social media to understand what customers and potential customers are saying about them and to identify trends and opportunities. Identifying fraudulent claims is another example. Here, the text written by claimants or other third parties is analyzed using text analytics. The extracted information is merged together with structured data typically found in claims forms. This combined information is integrated and used as part of a predictive model to gain greater insight into potentially fraudulent claims. Claims with a high score are sent to the insurance company's special investigation unit to examine more closely.

NUMBER SEVEN

EVALUATE THE AVAILABLE SKILLS AND CULTURE.

Technology generally provides little value if the skills and culture aren't available to utilize it to its fullest.

- **An imperfect science.** Working with text data is a bit different from working with structured data. It's often an imperfect science. For example, there can be trade-offs in text analysis between precision and recall. Precision refers to the fraction of instances that are relevant—with the goal to minimize incorrect information. Some cases require the highest precision possible—for example, if you're working with stock trades. If you're dealing with a legal discovery process, you must discover everything that matches, so precision is less important than recall. Often it is up to the analyst to tweak the system to fit the particular need. Sometimes being directionally correct is good enough.
- **Skills needed.** The skills needed to perform text analytics are actually similar to those needed for analytics in general (with the inexact science caveat noted above). These include critical thinking, understanding the data, a disciplined approach to problem solving, curiosity, communication skills, and the need to be able to defend the analysis. The level of actual analytical skills depends on the problem that you're trying to solve. Generally, understanding natural language processing is not a prerequisite for text analytics, although it is best to assume that some training on the text analytics tool will be necessary. Depending on the sophistication of the tool, someone may need to learn a scripting language. If you're including predictive analytics or other kinds of advanced analysis, the analyst will need to understand those, as well.
- **Cultural issues.** Change can be hard. Cultural issues with text analytics generally include managing initial skepticism as well as managing expectations. Communication is critical to get buy-in and sell the vision. A proof of concept can help you demonstrate value. Once a deployment is in place, keep the results front and center. For example, some companies share results frequently, using verbatim comments from text sources to get attention and keep the analysis current.

NUMBER EIGHT

WEIGH TECHNICAL FEATURES.

There are many technical features to consider when selecting a text analytics solution. In addition to those already mentioned, you may need high-performance text mining for large document collections, or features such as document summarization, document de-duplication, and metadata management. Here are some additional considerations:

- **Taxonomies and lists.** A taxonomy is a method for organizing information into hierarchical relationships. Taxonomies are used as part of features such as concepts and entities. For example, a telecommunications service provider might offer wired and wireless service. Within the wireless service, the company may support cellular phones and Internet access. Within the cellular phone service, there might be multiple models of phones. The taxonomy is then used to configure the different levels of classification that you want to see. Maybe you don't care about the different models of cell phone and you just want to get down to the granularity of "cell phone," or maybe you need to be able to split out all the different models. This configuration will generally happen in a list or taxonomy. Sometimes vendors offer industry taxonomies out of the box, but these usually need to be customized. Users should expect that no taxonomy will work perfectly out of the box. Several iterations will be needed to build a useful taxonomy.
- **Multi-lingual support.** Text comes in many languages. Vendors offer different solutions to the language issue. Some offer native language extraction (i.e., the ability to actually understand a foreign language). Others provide language translation capabilities. You must decide what works best in your particular situation. Note that there are nuances in languages that can make machine translation unreliable, particularly for sentiment analysis.
- **Speed and scalability.** When processing large content sets or time-sensitive content, the system should scale across multiple processors and work quickly for each piece of content. Text analytics is generally about an order of magnitude slower than a simple search indexing process.

NUMBER NINE

IMPLEMENT, TUNE, TEST, AND ITERATE.

No text analysis system is completely optimized when it first boots up. Most systems are tested and updated regularly, with results improving each time. This is necessary for a number of reasons, including the introduction of newly coined terms into a corpus as well as changes to the ways people use language (often called language drift). It is important to plan for this when you're past the proof-of-concept stage.

Take time to regularly test and tune your system to improve its accuracy. This can be done manually (either by doing it yourself or by using a service such as Amazon Mechanical Turk). Once you understand where your system isn't performing well, you can use the vendor-provided tools (such as machine learning) to help configure or retrain your system.

In a *supervised* approach, a training corpus (about half the corpus) is used to generate rules that include the presence or absence of terms needed to define a concept. A corpus of, say, 2,000 pieces of content might be used for training purposes. The model is then tuned to improve accuracy. A validation sample (the rest of the corpus) is utilized to test the accuracy of the model and to ensure that you didn't over-fit it to the first set of content. This is a real consideration when it comes to big data where you might want to model everything. Once the model is in production, thresholds are defined and monitored for drift. Finally, the model is retrained if necessary.

Machine accuracy is directly tied to the quality of the content. One best practice is to evaluate the level of agreement among several quality ratings performed by humans. Note that even the best-trained humans, working with very clean content, will agree only about 80% of the time when they perform a task such as evaluating the sentiment of a sentence. If agreement among these ratings is low, your natural language processing system will be hard pressed to achieve excellent accuracy. By following a process as outlined above, on content where your raters have good agreement, you can nudge your machines' accuracy higher.

ABOUT OUR SPONSORS

www.angoss.com

Angoss is a global leader in delivering predictive analytics to businesses looking to improve performance across sales, marketing, and risk. With a suite of desktop, client-server, and big data analytics software products and cloud solutions, Angoss delivers powerful approaches to turn information into actionable business decisions and competitive advantage.

Angoss software products and solutions have gained broad user acceptance—including 9 of the top 15 global financial institutions. They are user-friendly and agile, making predictive analytics accessible and easy to use for technical and business users alike.

Many of the world's leading financial services, insurance, retail, telecommunications and information communication, and technology organizations use Angoss predictive analytics software products and solutions to grow revenue, increase sales productivity, and improve marketing effectiveness while reducing risk and cost.

Angoss focuses on helping businesses leverage their data to discover the key drivers of behavior, predict future trends and outcomes, and act with confidence when making business decisions. The company's innovative software products and solutions are ideal for businesses looking to gain a competitive edge and improve profitability by putting their business data to work.

Headquartered in Toronto, Canada, Angoss has offices in the U.S. and U.K.

ABOUT OUR SPONSORS

www.lexalytics.com

A text mining company founded in 2003, Lexalytics was:

- First with sentiment analysis
- First with multi-level sentiment
- First with automatic topic detection
- First with integrating Wikipedia as
- a knowledge base

Lexalytics' software for text analysis, Salience, is engineered for easy integration into third-party applications, and is a critical component in many content processing services and applications for industries such as:

- (Social) Media Monitoring
- Voice of Customer
- Reputation Management
- Online Media
- eDiscovery
- Cyber-Intelligence

And lots more!

Our customers process over 3 billion documents per day through our text mining engine, looking for named entities, discovering topics, summarizing, and ascertaining sentiment.

In other words, we tell our customers "who", "what", "where", "when", and "how", so that they can help their customers figure out "why".

Read Between the Lines™ with Lexalytics.

www.sas.com

Who is SAS? When you think SAS, you think analytics. That's because SAS was the first provider of advanced analytics software, and after more than 35 years, we remain the market leader. We combine the strengths of our business solutions and technology infrastructure so customers have the ultimate flexibility in using analytics to solve their specific business problems.

How do we deliver these benefits?

- **Analytics.** We are the world leader in advanced analytics that fuel evidence-based answers. With our analytics, organizations can measure what matters today, reveal best actions, expose threats, and incorporate learning into business processes.
- **Data management.** We help customers manage a deluge of data that is quickly increasing in volume, velocity, frequency, and complexity. We offer a unified approach that includes data integration, data quality, data mastering, and enterprise data access.
- **Business intelligence.** Our software makes fact-based decisions possible by providing the right information to the right people at the right time. Our intuitive interfaces allow everyone—at every skill level—to access and produce reports quickly.

Simply stated, we give our customers THE POWER TO KNOW®

ABOUT THE AUTHOR

Fern Halper, Ph.D., is director of TDWI Research for advanced analytics, focusing on predictive analytics, social media analysis, text analytics, cloud computing, and other "big data" analytics approaches. She has more than 20 years of experience in data and business analysis, and has published numerous articles on data mining and information technology. Halper is co-author of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, service management, and big data. She has been a partner at industry analyst firm Hurwitz & Associates and a lead analyst for Bell Laboratories. Her Ph.D. is from Texas A&M University. You can reach her at fhalper@tdwi.org, or follow her on Twitter: @fhalper.

ABOUT TDWI RESEARCH

TDWI Research provides research and advice for business intelligence and data warehousing professionals worldwide. TDWI Research focuses exclusively on BI/DW issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence and data warehousing solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

ABOUT THE TDWI CHECKLIST REPORT SERIES

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.