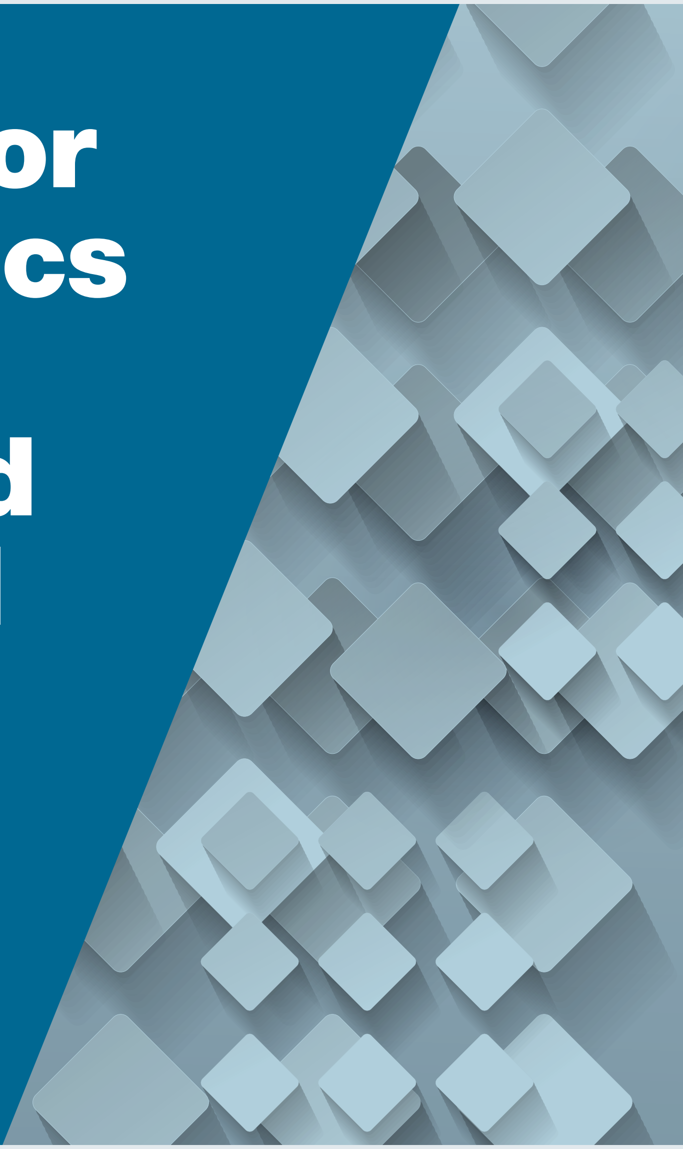
TDWI RESEARCH

TDWI CHECKLIST REPORT

Eight Steps for Using Analytics to Gain Value from Text and Unstructured Content



By David Stodder

Sponsored by:

tdwi.org

MAY 2014

TDWI CHECKLIST REPORT

EIGHT STEPS FOR USING ANALYTICS TO GAIN VALUE FROM TEXT AND UNSTRUCTURED CONTENT

By David Stodder

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

TABLE OF CONTENTS

FOREWORD

Organizations that aspire to be data driven cannot afford to limit users' data analysis to structured data: that is, the alphanumeric data types that have been carefully defined, modeled, and stored in standard spreadsheets, relational databases, and data warehouses. In the course of business processes, operations, and customer interactions, enterprises generate far more than just structured data—in particular, most are swimming in massive amounts of text.

Drawn from both internal and external sources, text files form a potential gold mine of insights that could help organizations reduce costs, improve customer relationships, speed response to events, and innovate with products and services. Business intelligence (BI) and data warehousing (DW) systems are adept at delivering the numbers, but if users do not have access to unstructured and semi-structured text such as customer comments, field personnel notes, and social media, they will be blind to contextual information that could help answer questions about the numbers. Exploratory analysis of text could reveal emerging trends that do not show up in BI reports.

Text analytics covers a range of technologies and practices for analyzing text, extracting relevant information, and transforming sources by applying structure so that analysis can be repeated and adjusted over time. Software solutions combine techniques from natural language processing, statistics, and machine learning. One of the goals of text analytics is to accurately extract entities, facts, concepts, themes, and sentiment. With the popularity of social media, many enterprises are interested in performing sentiment analysis. However, interest is expanding across industries to other uses, such as in healthcare, where text analytics helps doctors understand the full context of patient symptoms and engage in evidence-based medicine.

Making use of modern data visualization, text analytics solutions enable users to explore text on desktops and mobile devices. Business users, not just specialized text analysts, can use the solutions to derive value—for example, by aggregating findings from text with their analysis of structured data. Following TDWI's previous Checklist Report on this topic, *How to Gain Insight from Text*, this Checklist discusses the relevance of text analytics for improving user experiences with the semi-structured and unstructured textual big data now at their disposal.

NUMBER ONE

IDENTIFY BUSINESS PROCESSES AND OPERATIONS THAT WILL BENEFIT FROM TEXT ANALYTICS

Business processes often suffer from poor information awareness. Marketing teams may be launching campaigns that are less effective because of incomplete views of what customers or prospects desire or how they prefer to be engaged. Efforts to fight fraud are hampered by an enterprise's inability to relate information about activities such as illegal claims made at different touch points. Regulatory compliance processes fall short because of difficulties in analyzing complex policies, leading to metrics based on the wrong indicators. Most enterprises could cite numerous other examples affecting their operations.

Text analytics can play an important part in enabling organizations to optimize business processes to be smarter, more efficient, and more effective. In particular, text analytics can help those accountable for operational tasks and decisions to expand their view beyond structured data to the volumes of information generated by human interactions that occur during the course of business, which are most often captured in text. Without this fuller view, managers and frontline personnel will invariably make mistakes. Processes will miss the mark, adding costs and leading to misallocation of resources.

TDWI Research finds that the growth in semi-structured and unstructured text is overwhelming operational decision makers at both management and frontline levels. Operational intelligence, which focuses on the management and optimization of daily business operations, benefits from the ability to analyze text as it is being collected during customer or other human interactions. Frontline users could share such analysis with managers; text-based insights could be aggregated with structured or other data, such as photos. Text analytics can help reduce delays in examining call center records, warranty claims, and other text generated by events such as product purchases, shipments, and damage claims.

At this step, organizations should look at their business processes and operations to spot where a fuller view of text could play a major role in increasing efficiency and effectiveness. Often the best place to start is where human interactions generate large volumes of text, such as in customer service and call center processes. Along with improving customer satisfaction and loyalty, implementation of text analytics can reduce costs by making critical information and analysis more easily accessible.

☑ NUMBER TWO

ADDRESS BUSINESS SCENARIOS WITH THE RIGHT COMBINATION OF TEXT ANALYTICS TECHNIQUES

Text is everywhere. Sources may contain millions, if not billions, of data points about customers, products, and other topics of interest. As size and complexity increase, so will the cost of retrieving, analyzing, and classifying the information. To reduce costs and accelerate results, organizations must evaluate the advantages of quantitative analysis techniques and software automation over manual methods.

Text analytics projects usually involve several techniques. Different business scenarios will demand different assortments of tools, algorithms, and practices; improving content search could be the goal for some scenarios; automating classification of call center notes would be appropriate for others. What's critical is to match requirements with the right combination, which could include natural language processing, search, information retrieval, entity extraction, computational linguistics, classification, visualization, and more.

To enable efficient retrieval and analysis, the initial goal could be to organize and categorize content. Records such as doctor's notes, case histories, and repair calls can be so chaotic that the first job for text analytics is to solve the "variety" problem through classification. Text analytics solutions can either discover structure or enable you to develop a structure so you can understand what's buried in the content, then organize the data sourced into the structure for analysis and reporting. Classification, including development of taxonomies, is often critical; through teaching the software, systems learn to automatically recognize combinations of words that fit classifications and make distinctions based on meaning. Clustering programs discover groupings among people or other objects based on different types of models for defining what constitutes a cluster.

Some tools provide prebuilt taxonomies, the functionality for discovering them from document collections, and the ability to crawl external sources such as social media, tagging the Web material as it is being sourced. Some tools support development of ontologies, which provide higher-level conceptual views above single taxonomies. Text analytics can use taxonomies and ontologies to perform deeper analysis.

The focus at this step is to address business scenarios with a strategy based on the multipurpose nature of text analytics and classification approaches. Evaluate how software can help you deal with the magnitude and diversity of sources and avoid costs through automation.

☑ NUMBER THREE

DEVELOP A STRATEGY FOR ANALYZING UNSTRUCTURED DATA FROM SOCIAL MEDIA

Social media is where customers, prospects, and influencers freely express themselves. Their data trails are a potential gold mine, which in recent years has made social media data a major data source for text analytics. Organizations want to mine social media and aggregate findings with analysis of internal customer and market data.

Marketing functions can analyze social media text to discover how people regard their company's brand, products, and services. In addition to being sources of data, social media networks form a critical channel for marketing communications; thus, what the organization learns from analyzing text data generated by social media activity can inform marketing strategies in real time to guide ongoing messaging and interaction.

Usually, the focus is to analyze sentiment. Most sentiment analysis processes are aimed at determining the "polarity" of sentiments expressed in social media. Many websites allow visitors to rate something and provide comments to accompany the rating. Often, the comments provide the most revealing data, adding color and opinions that can be quite different from the structured data recorded in surveys.

Text analytics enables organizations to interpret the sentiment in comments on a large scale. Algorithms can be tuned to filter out the noise and evaluate meaning, which can be confusing because people will use sarcasm, double negatives, and diverse colloquial expressions. Text analytics can apply both statistical and linguistic techniques for sentiment analysis. For many projects, the aim is to develop rules for identifying sentiment and then score documented feedback, such as comments made after product introductions.

Visualization helps decision makers understand the significance of sentiment analysis results. It can be as simple as bar charts with red indicating the degree of negative sentiment and green for positive. Network graphs help users see and evaluate the impact of social media "influencers" who shape others' decisions about products, services, and brands.

In this step, organizations should evaluate visual text analytics for interpreting and exploring social media sentiment. In general, sentiment analysis is an inexact science; organizations should set expectations carefully to ensure that users apply insights wisely.

NUMBER FOUR

DEPLOY TEXT ANALYTICS TO GAIN VALUABLE INSIGHTS FROM CUSTOMER INTERACTIONS

The analyst firm Gartner projected that by 2017, chief marketing officers will outspend CIOs on technology purchases. Gartner has said that high-tech marketing budgets have been growing at more than twice the rate of IT budgets, in part due to the competitive need to analyze big data. Because the lion's share of big data is text, it's clear that over the next few years, many organizations will focus on how they can cost-effectively use software solutions to improve customer analytics against text sources. In addition to strategic uses, enterprises will use text analytics to make automated interactions such as live chat smarter and more valuable.

Customer touch points today include e-commerce and social media, along with traditional avenues such as call centers, field sales and service, and physical stores. Conversations are occurring at all touch points, creating voluminous text files. Organizations need to listen and continuously learn from customer interactions. This can be difficult because the information recorded in online chat, field notes, e-mail, and message boards can be noisy: that is, full of errors, repetitions, and unusual syntax, among other problems. Text analytics can apply statistical and linguistic methods to reduce the noise and deliver results tailored to the scope of analysis.

Leading organizations are using text analytics systems to aggregate multiple unstructured data types with other sources, such as a customer data warehouse containing transaction notes and demographic data. Many text analytics tools can turn raw data collected from a range of content sources into structured information to support implementation of predictive text analytics. Predictive insights help organizations be proactive; with the insights, they can anticipate changes in customer preferences and act in advance of competitors.

Marketing processes where text analytics has proven useful include problem resolution, churn reduction, segmentation, brand management, and improvement in lift from marketing programs based on analytical models. Some organizations additionally use text analytics to find specific words in customer interactions that they can use effectively in corporate speeches, promotions, and customer communications.

At this point, organizations should evaluate where in their interaction processes text analytics could provide faster insight, particularly to support more intelligent automation.

NUMBER FIVE

PLAN TO INTEGRATE ANALYTICS FOR STRUCTURED AND UNSTRUCTURED DATA

The worlds of structured data and unstructured or semi-structured content have historically been separate. Structured data has been managed and analyzed for business intelligence querying and reporting, while text and other content sources have served processes that include Web publishing, document sharing, meeting legal and regulatory needs, and other search and information retrieval requirements.

Now, however, these worlds are colliding as users seek greater information access. In dashboards and portals, users want to track, analyze, and visualize information about objects of interest drawn from a variety of data types. TDWI Research finds that the appetite for unifying information access is strong, and that tools will continue to evolve toward greater integration of functionality for search, querying, classification, and analytics.

Here are three important objectives that TDWI Research finds are driving greater unification:

- **Single views of all information**. Disconnected information silos prevent organizations from gaining complete views of customers, supply chains, fraudulent activity, business performance, and more. Decision makers can benefit from a single view that fills in the gaps and reveals relationships across information sources.
- **Fuller context for structured business data**. Structured data listing frequency and codes from records of transactions, equipment orders, healthcare diagnoses, and more are critical for analysis. However, reports based on just these sources do little to help users answer the "why" questions: Why are these numbers so high (or low)? By providing access to information generated by subject matter experts' notes or customer interactions surrounding data transactions, users can gain a faster path to insight.
- **Better attributes for predicting behavior**. Analysts building predictive models to understand and anticipate customer churn, fraudulent claims, payment delays, or a rise in service calls require the best possible mix of attributes. Attributes generated by social media data, contact center notes, field service notes, and more could be combined with those from structured sources to gain new predictive insights.

At this step, organizations should evaluate whether their current user workspaces and analytics processes can provide integrated access to structured, semi-structured, and unstructured data. Organizations should plan to give users an integrated view to relieve them of the drawbacks of working in disconnected data environments.

☑ NUMBER SIX

INCREASE DATA VISUALIZATION OPTIONS TO ENABLE CLEARER INSIGHT FROM TEXT SOURCES

Graphical visualization and interaction with data is now the expected norm for most users, including executives, departmental managers, and an increasing number of frontline personnel in sales, service, inventory, and other functions. No matter what data types are used as sources, good visualization is critical to making smarter decisions and improving user productivity by providing a faster path to insight. In addition, because pictures are easier to consume, visualization can foster better collaboration on data among users in different roles and functions.

Visualization is a key part of the trend toward self-service functionality in BI and analytics discovery. Tools are making it simpler for users to select visualizations such as charts, data maps, heat maps, histograms, and scatterplots to suit their analysis, reporting, and data sharing needs. Dashboards can facilitate a mashup of views that employ several types of visualization (see Figure 1). Many users want to extend this functionality to text-based content so they can visualize word frequencies, patterns, and trends across all available types of data.

Many users begin visualizing text with word clouds, which show the relative importance of terms or concepts as they appear across all text sources. Word clouds can be helpful for understanding social media trends, term and theme trends, as well as expressions used in other documents, such as call center notes. Filtering can help users reduce noise and interact with specific slices of data to focus queries. Some tools offer functionality beyond word clouds, including network diagrams and term constellations (see Figure 2). These enable users to more easily visualize relationships, including the use of words or concepts in different geographies.

As visualization becomes more common for all types of data, organizations need to pay attention to the quality and relevance of the presentation and set user expectations accordingly. Poor, static visualizations that can't be explored or deliver misleading views can lead to poor decisions and a lack of trust. Organizations should also tailor visualizations to users' roles and responsibilities so they have actionable information that supports their decision-making needs and the actions they take based on the information.

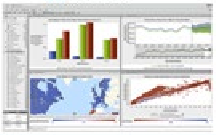


Figure 1

Click to enlarge

Figure 2

Click to enlarge

☑ NUMBER SEVEN

DEPLOY TEXT ANALYTICS WITH HADOOP, IN-MEMORY COMPUTING, AND OTHER BIG DATA TECHNOLOGIES

As mentioned earlier, the big data trend is dominated by a desire to tap growing volumes of content for business insights. Enterprises need a fuller view of information beyond the confines of structured data. Executives and marketing management, chief data officers, and line-of-business management are major drivers of big data technology deployment, according to TDWI Research. Performing analytics on marketing and social media data is high on the list, but other priorities include competitive intelligence, fraud detection, risk management, and financial management.

The rising interest makes it clear that big data is hardly just about data volume or the largest organizations with the biggest databases. Small and midsize organizations also struggle with how to integrate and analyze their variety of data, which can include records of customer interactions, service call notes, documents, books, forms, and e-mail messages. Along with reporting, enterprises need to perform discovery analysis, where they look for patterns and previously unknown trends.

Text analytics helps businesses develop classification structures in big data and extract entities, themes, concepts, facts, and sentiment to filter out the noise. Some tools employ statistical models based on machine learning to speed extraction from big data "lakes" where there is no structure. Today, these sources of raw data are often held in Hadoop files. Organizations can use Hadoop and related technologies such as MapReduce to support implementation of sophisticated text analytics algorithms and models across raw, detailed data to find patterns, unearth correlations, and spot anomalies that could be important to their objectives. Using data visualization, business users can examine the results of analysis and iterate through data to answer questions more easily.

Demand for speedier insights is increasing the importance of in-memory computing to provide the power and scale needed for visual text analytics on big data sources. In-memory computing is an alternative to reading data from disk for every operation. It can allow organizations to bring larger volumes of data closer to users so they can perform self-service analytics. Organizations should evaluate the role of in-memory computing to support text analysis of big data.

☑ **NUMBER EIGHT**

ENABLE MOBILE DEVICE USERS TO BENEFIT FROM VISUAL TEXT ANALYTICS

Mobility continues to be an important trend. On the go with smartphones and tablets, users want the latest devices that allow them to download apps, personalize them, and share content with others via texting, social media, and other methods. Traditional BI systems that have focused on providing users with standard reporting and analysis of only structured data are under pressure in an environment where users want more types of data, better visualization, and self-service options tuned to the mobile experience.

Initially, many firms were hesitant about mobile BI and analytics because of security concerns. Security is still a key issue, but with experience, leading organizations are developing strategies for managing data. As users seek access to semi-structured and unstructured data sources for text analytics, security processes must be revised. Enterprises also need to set user expectations for performance, data quality, and availability for text analytics just as they have for mobile access to traditional data. Organizations must consider the context of users' mobile experience—that is, whether they are on site with customers, at a plant or subsidiary, or in the midst of collaboration with other users—to tailor performance, availability, and functionality.

Mobile text analytics can be beneficial to many types of users, such as field sales and service managers who need to explore context around structured data. Managers at the home office, for example, could visualize trends found in service notes and claims and share their insights with field service personnel currently engaged with accounts.

Medical health professionals could also benefit by using text analytics to look for word frequency and other patterns in text messages; with these insights, they could fine-tune treatments or recommend changes in patient behavior. As evidence-based medicine becomes standard, healthcare professionals will need stronger capabilities for accessing and analyzing text sources while engaged with patients.

Organizations should not ignore the powerful trend toward mobile access and analysis. At this step, develop a strategy for enabling visual text analytics on mobile devices and initiate data governance procedures to accommodate mobile access.

ABOUT OUR SPONSOR

www.sas.com

Organizations of all types and sizes constantly generate both structured and unstructured data. Everyone—from executives to employees on production lines—needs data-driven insights to improve decisions and take smarter actions.

SAS Visual Analytics enables you to explore all relevant data easily, in minutes or seconds—no matter the size of your organization or data. Anyone—regardless of analytic background—can benefit from complex data, because sophisticated analytics have been seamlessly integrated.

Explore and share data with compelling visuals. Create forecasts and hierarchies on the fly. Use auto charting to automatically select the best visualizations for your data. Quickly understand relationships and patterns with "what does it mean" pop-ups. And that's just scratching the surface.

SAS Mobile BI is one of the critical components of SAS Visual Analytics, rendering reports and visualizations that can be shared with iPad and Android mobile devices. This provides rapid insights on the go—delivering secure, interactive information anytime, anywhere.

Seeing is believing. Visit sas.com/visualanalytics to test-drive the software, view reports and videos, and explore data.

SAS is the leader in business analytics software and services, and the largest independent vendor in the business intelligence market. Since 1976, SAS has been giving customers around the world THE POWER TO KNOW.

ABOUT THE AUTHOR

David Stodder is director of TDWI Research for business intelligence. He focuses on providing research-based insight and best practices for organizations implementing BI, analytics, performance management, data discovery, data visualization, and related technologies and methods. He is the author of TDWI Best Practices Reports on mobile BI and customer analytics in the age of social media, as well as TDWI Checklist Reports on data discovery and information management. He has chaired TDWI conferences on BI agility and big data analytics. Stodder has provided thought leadership on BI, information management, and IT management for over two decades. He has served as vice president and research director with Ventana Research, and he was the founding chief editor of *Intelligent Enterprise*, where he served as editorial director for nine years. You can reach him at dstodder@tdwi.org.

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.