

MIS QUARTERLY EXECUTIVE

Using Text Analytics to Derive Customer Service Management Benefits from Unstructured Data

Deriving value from structured data is now commonplace. The value of unstructured textual data, however, remains mostly untapped and often unrecognized. This article describes the text analytics journeys of three organizations in the customer service management area. Based on their experiences, we provide four lessons that can guide other organizations as they embark on their text analytics journeys.¹

Oliver Müller

IT University of Copenhagen
(Denmark)

Iris Junglas

Florida State University
(U.S.)

Stefan Debortoli

University of Liechtenstein
(Liechtenstein)

Jan vom Brocke

University of Liechtenstein
(Liechtenstein)

The Growth of Text Analytics

Estimates suggest that about 80% of today's enterprise data is unstructured.² Unlike structured data, which is tidy and mostly numeric, unstructured data is often textual and, therefore, messy. Unstructured data comprises documents, emails, instant messages or user posts and comments on social media, and presents a challenge to data miners; analyzing unstructured data is more complex, more ambivalent and more time consuming. Extracting knowledge from unstructured text, also known as text mining or text analytics, has been, and still is, limited by the ability of computers to understand the meaning of human language.³ The written word, however, can provide valuable insights about the inner workings and environment of an organization; it has the potential to improve an organization's productivity while generating value for customers.

Some organizations are successfully leveraging textual data as part of their analytics efforts. For example, pharmaceutical companies mine patents and scientific literature to improve product development processes. Doctors use text analytics to aid medical diagnosis by mining electronic health records. Insurance companies analyze claims and damage reports to mitigate risk or detect fraud.



¹ Federico Pigni is the accepting senior editor for this article..

² See, for example, "Discover the Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," *EMC Digital Universe Study*, EMC (with IDC), 2014, available at <http://www.emc.com/leadership/digital-universe/index.htm#2014>; and Dhar, V. "Data Science and Prediction," *Communications of the ACM* (56:12), 2013, pp. 64-73.

³ See Jurafsky, D. and James, H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Pearson, 2009.

Market forecasters have estimated that, by 2020, the market for text analytics will reach \$5.93 billion.⁴ Today, 30% of business analytics projects include the delivery of insights based on textual data. In some industries (e.g., banking, retail, healthcare), growth rates for text analytics of up to 50% are not uncommon.⁵ Within organizations, text analytics is predominantly used by marketing and customer service units. Analyzing customer-related data in the form, for example, of customer call notes or service requests, allows organizations to better understand their customers' problems, react to service issues in a timely fashion and proactively plan process improvements that can alleviate problems in the long term. For instance, by mining the stream of service requests flowing into and within an organization, enterprises can identify the issues of greatest concern to their customers, monitor how issues evolve over time, "slice and dice" issues by customer groups, product categories or other dimensions and recommend solutions for recurring issues.

This article examines three organizations that aggressively use text analytics to improve customer service management. All three use text mining to analyze the content streams of incoming service requests so they can better understand their customers' problems and, as a consequence, improve their customer service processes.

The three organizations come from different geographies (U.S. and Europe) and industries (research and education, manufacturing and finance), provide different types of services (IT services, and maintenance and repair of physical goods) and serve different types of customers (internal and external). All three are faced with similar challenges in harnessing the enormous amounts of unstructured textual customer feedback they receive during service interactions.

4 *Text Analytics Market by Applications (Marketing & Customer Experience Management, Data Analysis & Forecasting, Enterprise Information Management, & Other Industry Specific Applications), Deployment, Vertical, & by Region - Global Forecast to 2020*, marketsandmarkets.com, 2015, available at <http://www.marketsandmarkets.com/PressReleases/text-analytics.asp>.

5 Duncan, A. D., Linden, A., Koehler-Kruener, H., Zaidi, E. and Vashisth, S. *Market Guide to Text Analytics*, Gartner, 2015, available at <https://www.gartner.com/doc/3178917/market-guide-text-analytics>.

A New Generation of Text Analytics Solutions

The goal of enabling computers to understand natural language is as old as commercial computing itself.⁶ Since the early 1950s, researchers have been trying to automate the analysis of human language—with mixed success. It wasn't until the 2000s when the rise of statistical machine learning approaches, driven by the increasing availability of digitized texts from the World Wide Web and by increases in computing speed and memory, made text analytics feasible for commercial applications.

Early natural language processing systems were developed by and for computational linguists and were well beyond the understanding of the average business analyst. In contrast, the latest generation of text analytics solutions is more usable. These solutions possess four distinctive features.

First, advanced text analytics tools are *data-driven* rather than rule based. While early natural language processing systems were based on hand-written grammatical rules for extracting meaning from texts, today's systems rely on statistical machine learning techniques to automatically discover patterns in large collections of texts. As a result, they can process a wide variety of texts from different domains without needing laborious updates to an underlying rule base. They are also able to handle ungrammatical texts—for example, posts taken from online forums or messages from social networking sites. In addition, the inductive nature of today's solutions ensures that they learn over time—the more data they crunch, the better the results. For example, one of the key success factors of IBM's Watson, a supercomputer that beat two former (human) grand champions on the game show "Jeopardy!," is its ability to integrate new data sources with minimal efforts and to increase its performance with each new data source added (e.g., books, dictionaries, webpages, Wikipedia).

Second, the latest text analytics systems are able to process textual data in *real time* rather than in periodic batch runs. Increases in

6 See Jurafsky, D. and James, H., op. cit., 2009; and Manning, C. D. and Schütze, H. *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.

computing power and storage capacities make it feasible to analyze textual data as it streams into an organization. For example, the Twitter application programming interface (API) provides access to more than 6,000 tweets a second, or 500 million tweets per day—a feature that Starbucks uses to “listen to the voice of its customers.”⁷ During new product launches, Starbucks performs sentiment analyses of Twitter and other blogs and discussion forums to determine the success of a new product in near real time and to react to potential issues on the spot (e.g., by reducing prices or changing the coffee blend).

Third, today’s text analytics solutions make *probabilistic* inferences rather than drawing deterministic conclusions. While rule-based systems by nature provide a single answer, probabilistic systems provide alternative answers to any given query, each weighted with a likelihood of being correct or relevant. For example, IBM’s Watson generates hundreds of potential answers (hypotheses) to any given query and then weights each with a confidence score. Based on this confidence score, a human might decide to place sufficient trust in a machine-generated solution or ignore it. This can greatly increase the level of trust in situations

⁷ Watson, H. J. “Tutorial: Big Data Analytics: Concepts, Technologies, and Applications,” *Communication of the Association for Information Systems* (34), 2014.

where the stakes are high—for example, when trying to spot fake insurance claims or performing medical diagnoses.

Fourth, until recently the outputs produced by text analytics systems largely consisted of cryptic codes, attached as annotations to the original text. These outputs were primarily meant to be read and interpreted by a computer scientist or to serve as input for other algorithms. The latest generation of text analytics systems, in contrast, aggressively uses *visual displays* that communicate results in an intuitive and effective way. Twitter, for example, offers various interactive graphics (<http://interactive.twitter.com>) that enable non-specialists to explore and interpret the contents of the stream of tweets produced in reaction to various social, economic or political events. The visual outputs of text analytics tools, combined with the increased overall usability of the tools, enable ordinary and non-technical line-of-business people to analyze written texts themselves.

These four features of modern text analytics solutions are summarized in Table 1.

Each of the organizations discussed below used the same tool, which had these four features of advanced text analytics solutions. The tool was developed as part of the research effort on which this article is based and then adapted and tailored

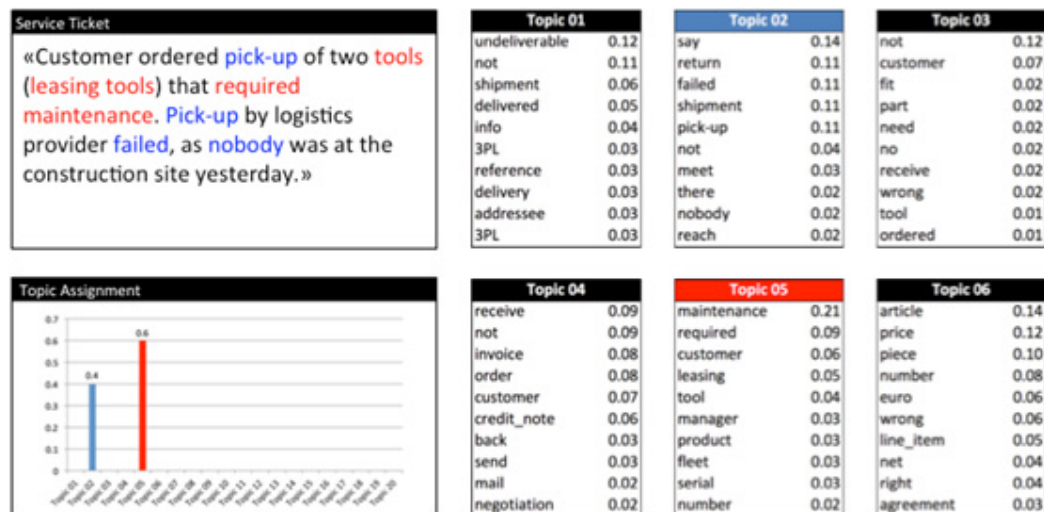
Table 1: Features of the Latest Generation of Text Analytics Solutions

	Feature	Business Value
Foundation	Data-driven instead of rule-based	<ul style="list-style-type: none"> Requires less manual effort Provides self-learning systems
Speed	Real-time instead of periodic batch runs	<ul style="list-style-type: none"> Enables continuous “listening” to textual data streams Reduces reaction times
Logic	Probabilistic inferences instead of deterministic decisions	<ul style="list-style-type: none"> Explores different answers to a given problem Gains trust of users
Output	Intuitive visualizations instead of cryptic annotations	<ul style="list-style-type: none"> Provides fast and effective communication to end users Enables self-service text analytics for lines of business communication to end users

Box: Text Analytics with Probabilistic Topic Modeling using the Latent Dirichlet Allocation (LDA) Algorithm

The idea behind probabilistic topic modeling algorithms like Latent Dirichlet Allocation (LDA) is that texts exhibit multiple topics in different proportions.⁸ For example, the service ticket displayed in the top left corner of the figure below is reporting a failed pickup of a machine tool that should be sent to a repair center. On the right hand side of the figure are words associated with various topics. For example, there are words about return logistics (e.g., “return,” “failed,” “pick-up”) and about maintenance (e.g., “maintenance,” “required,” “tool”). The bar chart at the bottom left corner shows how the given service tickets blends two topics (Topics 2 and 5) by providing probabilities of topic occurrences (i.e., 60% of the ticket talks about maintenance and 40% about return logistics).

The primary advantage of LDA is its ability to automatically discover topics and their word distributions from large collections of texts and to annotate individual texts with topic labels. Grouping and aggregating the probabilistic topic assignments for thousands of service tickets enables, for example, a service manager to quickly gain an overview of current issues and to track their development over time.



to the needs of each organization. The core of the tool is the ability to analyze large volumes of text in a probabilistic, data-driven fashion; it does the analysis in near real time and presents the results in an easy-to-understand visual form. The tool is based on the probabilistic topic-modeling algorithm Latent Dirichlet Allocation (LDA), which is described in the box. A base version of the tool is freely available at MineMyText.com.

Case Organizations

The three service organizations that participated in our research varied in the type of service offered, their audience and their industry as well as their geography. Overall, we conducted more than 25 interviews and workshops with representatives of these organizations. The range of informants included business and IT people ranging from executive to operational levels.

⁸ For a more detailed description of LDA, see Blei, D. M. “Probabilistic topic models,” *Communications of the ACM* (55:4), 2012, pp. 77-84.

Information Technology Services at Florida State University

Information Technology Services (ITS) is the central IT organization for Florida State University (FSU). It provides IT support to 16 colleges and more than 110 centers, facilities, labs and institutes, serving 11,000 faculty and staff along with more than 40,000 students. It offers more than 100 IT services, including email, desktop support, file storage, software licensing, classroom and learning management system support and online training. ITS is also responsible for FSU's overall IT strategy, setting and enforcing standards, and managing and ensuring compliance—a vital need for a state institution.

As part of an enterprise-wide CRM installation in 2011, ITS decided to add a component for capturing and managing IT service requests across business units. This component had multiple goals. Its primary purpose was to capture the flood of incoming service requests into one system and to manage their execution. But it also built a knowledge base of “best practices” and thus supported IT personnel by providing an easy-to-use and transparent way of solving IT service problems.

Today, ITS's system captures, among other things, textual descriptions and manual classifications of IT problems (e.g., desktop, email, network), when they were reported, who reported them and which channels were used to report them (e.g., email, phone). The system also allows agents to add updates in the form of notes and a description of the solution once a case is resolved. Requests are routed to different work groups depending on their category, type or detail specified by a customer. Each work group has a queue manager who triages requests and assigns them to agents based on workload. At present, more than 100,000 service requests and their resolutions have been logged by the system since its inception in 2011.

To better understand the issues and challenges of the service delivery process, ITS prepares a weekly report as part of its analytical efforts. This report details, for example, the number of requests issued in the past week by the various business units, the number of open and closed requests, along with the type of channel used to

send a request and the customer that requested it. Overall, the report provides a snapshot of the IT support group's case load as well as how well the group is meeting response time guidelines. It thus provides a base for ITS to educate support agents on proper case management.

The report represents a valuable summary of the major elements of the stream of incoming service requests, namely the *who*, *when*, *where* and *how* of a service event. In contrast, insights into the *what* and *why* of a service request are captured in natural language within the textual fields of the service tickets.⁹ FSU uses text mining on these textual fields to interpret the what and why of service requests, thereby supplementing the findings drawn from the numbers-based report.

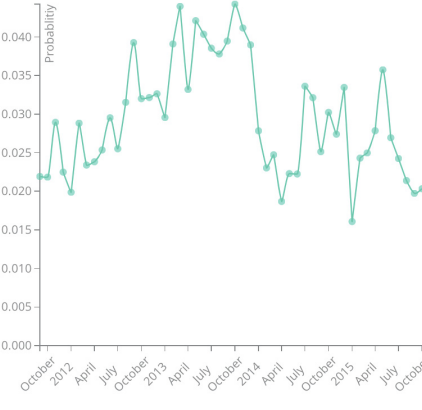
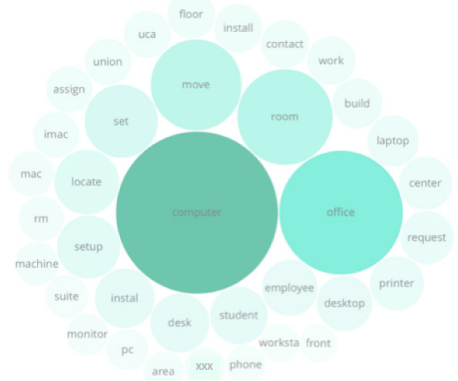
ITS applied “topic modeling” to more than 100,000 service tickets that had been received over the past five years. Figure 1 presents the seven most pressing topics to which ITS had to respond. Each topic is represented by a list of words that are associated with the topic (displayed visually as a word cloud), a sample sentence from a service ticket representative of the topic, a descriptive label for the topic and the prevalence of the topic plotted over time.

The identification of the most prevalent topics and their evolution over time provided ITS with a detailed picture of the history and evolution of service requests. For example, the analysis showed that requests to register new devices on the network (Topic 1) rose steadily in frequency from October 2011 (2.2% of all tickets) to October 2013 (4% of all tickets). This period coincided with an initiative to register all managed network devices by their unique MAC address for accountability and ease of management. In winter 2013, requests suddenly declined again, returning to previously observed levels (to less than 2%).

Other topics, like email forwarding (Topic 3), suddenly rose, which can be explained by an email platform migration that triggered new workflows and support processes at the time. Likewise, the rapid rise in phone services requests (Topic 4) can be explained by an ongoing

9 For more information on the elements of digital data streams, see Pigni, F., Piccoli, G. and Watson, R. “Digital Data Streams: Creating Value from the Real-time Flow of Big Data,” *California Management Review* (58:3), 2016.

Figure 1: ITS's Seven Most Pressing Service Requests Topics

Words describing the topic	Representative sentence from associated service tickets	Descriptive label	Timeline of topic prevalence
<p>Topic 1</p> 	<p>Good morning hostmaster. Please setup MAC assigned DHCP and update DNS as per the following information: host name: XXX*</p>	<p>Registering new devices on the network</p>	
<p>Topic 2</p> 	<p>Requesting XXX to set up an alternate log in on the Mac Computer located in Room XXX at Alumni Center Facilities. Computer belonged to former employee XXX. It will need to be backed up and re-imaged for a new employee whose potential start date is June 21st. Please contact me or XXX for further information.</p>	<p>Office move</p>	

<p>Topic 6</p>	<p>Need Building access and access to suite 3100 for two Graduate Assts. 1. XXX FSU CARD #: XXX 2. XXX FSU CARD #: XXX</p>	<p>Building access cards</p>	
<p>Topic 7</p>	<p>We have a computer that cannot access certain SharePoint features (data sheet view) due to having a 64-bit system. According to Microsoft support, this machine needs to download and install the 2007 Office System Driver: Data Connectivity Components. This installation requires an admin password. Please advise.</p>	<p>Software updates</p>	

*Due to the sensitive nature of the data, we have redacted portions of these entries with “XXX”

campus-wide initiative to move regular phone lines to a VOIP service.

The process of spotting and explaining historical service requests patterns raised ITS’s interest in analyzing more recent topics. For example, ITS was aware that installing and updating software (Topic 7) was a current issue. Textual analytics had confirmed how prevalent this topic was and had been over the years, providing additional empirical evidence for continuing an existing project aimed at streamlining software updates.

ITS also realized that the textual analysis of historical tickets validated the redefinition of its service request taxonomy that had been underway since 2013. The existing taxonomy used a three-tier architecture to classify service requests. When filling out an online request, a user was asked to classify the nature of the problem into one of 35 predefined categories. In the first tier, the user could choose, for example, between classroom and computer lab support, email, or logins, passwords and system access. The second and third tier (each comprising three

to five predefined categories) then allowed the user to specify their request even further. The text analytics findings provided a cross-validation for redefining the taxonomy as planned.

ITS's use of text analytics provided a glimpse into the possibility of extracting insights from the otherwise difficult-to-interpret unstructured elements of digital data streams. By using text analytics, ITS was not only able to supplement insights provided by the traditional quantitative report but also to validate efforts already underway.

Customer Service at Hilti

Hilti is a leading-edge technology provider to construction professionals, located in Schaan, Liechtenstein. Its many products, including power drills and demolition hammers, are the preferred tools used by the construction and building maintenance industry around the globe. Hilti markets its high-tech products through a direct sales model, with two-thirds of its 20,000 employees having daily customer contact involving more than 200,000 customer interactions. A large part of these interactions are handled by Hilti's multi-channel contact center, where agents capture each of the one million incoming service requests per year as a so-called customer care note in a CRM system and manage its resolution.

The global contact center is known for its fast response in handling individual customer issues. However, extracting strategic knowledge from the steadily expanding database of transcribed phone calls and incoming emails requires enormous efforts. Once a month, data is extracted from the CRM system and imported into a custom database for further analysis. The objective of this analysis is to unearth the root causes of recurring product failures, monitor persistent complaints about customer service and identify potential business process improvements. Findings from the analysis are discussed as part of a quarterly meeting of executives from all business units.

The analysis for the quarterly meetings is carried out manually, using a combination of searching, filtering and sorting through the dataset, which makes the findings susceptible to oversights. If, for example, the analyst uses a search term different from the terms used by the originator of a service request to

describe a problem, the findings can easily become distorted. This well-known "vocabulary problem"¹⁰ of text-based human-computer interaction is especially severe in large and heterogeneous domains where many different types of users interact with a system. Hilti's service requests, for example, include many technical terms and abbreviations, and are written by a heterogeneous group of users, ranging from service agents with their own varying backgrounds to customers from different industries and geographies.

Another problem that has made the manual analysis cumbersome is rooted in Hilti's service process. Agents are required to classify requests into just one of 20 categories, even if a request matches more than one category or an appropriate proper category does not exist. This forced classification, which is driven by a predefined classification system, not by the actual data, might lead to a biased picture of customer feedback and even prevent newly emerging issues from being detected in a timely manner.

The time lag between the occurrence of events and their analysis, combined with the largely manual analysis of patterns, meant that Hilti was missing out on many opportunities for capturing business value from the digital streams of customer feedback. As part of our research effort, we proposed that these deficiencies might be addressed by a probabilistic and data-driven analysis of customer care notes through the use of topic modeling. A team of six, including two from Hilti's CRM team, two from Hilti's IT team and two researchers, was formed to explore the feasibility of this approach. The first prototype was tested in Hilti's German market during spring 2014 and captured about 50,000 service requests over a six-month period. Later that year, a second prototype with a total of 30,000 service requests from Hilti's French market and spanning a period of four months, was built.

These prototypes were built around a dashboard that allowed the continuous inflow of service requests to be summarized and explored by topic. More specifically, the system visualized the overall distribution of topics, provided the ability to drill down into individual topics, traced

¹⁰ Furnas, G. W., Landauer, T. K., Gomez, L. M. and Dumais, S. T. "The Vocabulary Problem in Human-System Communication," *Communications of the ACM* (30:11), 1987, pp. 964-971.

Figure 2: The Tree Map of Topics is Part of Hilti’s Text Analytics Dashboard



each topic’s evolution over time and “sliced and diced” topics by geography or customer segments.

Figure 2 depicts the entry point into Hilti’s dashboard. It shows a tree map that visualizes the prevalence of a particular topic by the size of its rectangle and its semantic similarity with other topics by its position and color. This kind of visualization is based on the fact that some topics are similar to each other in meaning and sometimes even overlap. For example, Topics 1 and 13, displayed as light green in Figure 2, are both logistics related: one represents delivery failures, the other delayed deliveries. Likewise, orange signifies product-related issues (e.g., maintenance and repair of machine tools), and purple signifies topics related to invoicing.

By clicking on one of the topics in the tree map, users are led to a details page that provides further information about the topic. Among other things, this page shows the words that are associated with the topic, as well as its evolution over time. Figure 3, for instance, shows the details for Topic 5 (“Pick-up failed because of wrong address”). It includes a word cloud containing the most probable words associated with the topic, along with a stacked line chart representing the development of the topic over time, broken down by customer segments.

For example, the graph documents a rise in failed pick-ups for customers in the “E” segment (brown area). Drilling further into the data for the “E” segment and displaying affected customer names and locations (not shown in Figure 3) allows Hilti to identify potential root causes of the failures.

For the future, Hilti envisions enhancing the dashboard by providing automated alerts triggered when a topic varies by more than two standard deviations from the norm. These alerts will help Hilti to address emerging customer issues in near real time.

A major part of the text analytics project at Hilti was dedicated to evaluating and improving the accuracy of the topic-modeling and visualization algorithms. The solution passed through several iterative development cycles to identify the best possible set of parameters. These included determining the ideal granularity of topic modeling by increasing or decreasing the number of topics to be extracted or varying the natural language preprocessing pipeline to remove “noise” from the input data. Output from the solution was also repeatedly compared to categorizations made by human experts. Experiments showed that the solution agreed with expert judgments in 77% of cases, which

Hilti deemed to be more than satisfactory, given the time savings that could be realized by automating the text analysis process.

Hilti’s main goals in mining unstructured digital data streams were related to achieving efficiency gains and enhancing decision making in the area of customer service. More specifically, Hilti was interested in streamlining internal operations to save costs, while, at the same time, improving customer service. By leveraging text mining and visual analytics, Hilti increased its ability to interpret unstructured textual streams, which enabled it to monitor the feedback of thousands of customers in real time and provided the possibility of responding to patterns and trends in an ad hoc fashion.

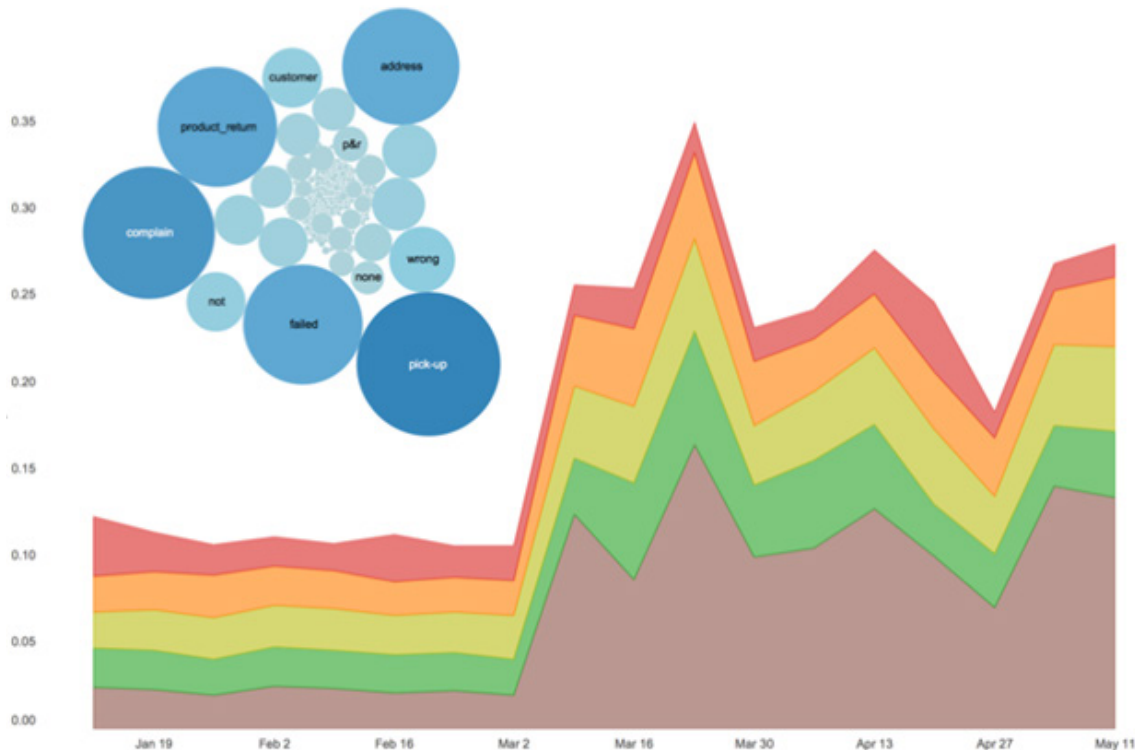
IT Service Outsourcing at Inventx

Inventx AG is a Swiss IT company that specializes in providing IT services to leading financial institutions. As well as providing IT service management solutions, its portfolio comprises the planning, implementation, integration, hosting and maintenance of core

banking solutions. In addition, Inventx runs one of the most modern data centers in Switzerland. A central part of its business model is providing outsourcing services to banks for back-, middle- and front-office IT applications. Ensuring continuous availability and high performance of these applications is vital both for Inventx and for its customers. To ensure superior service, Inventx operates its own helpdesk for managing and resolving IT service requests.

A major part of a service agent’s daily routine is driven by scripts that capture how a particular IT problem can be solved. In the past, these scripts were written by agents for agents, using Microsoft OneNote. OneNote is a collaboration tool that allows users to capture typed as well as handwritten notes, screenshots and audio snippets, and to share them with others. These scripts are written in a free format and edited by hand; their existence, accuracy and currency depend highly on the intrinsic motivation of agents, who volunteer to contribute their knowledge to OneNote.

Figure 3: Drilling Down into One Topic



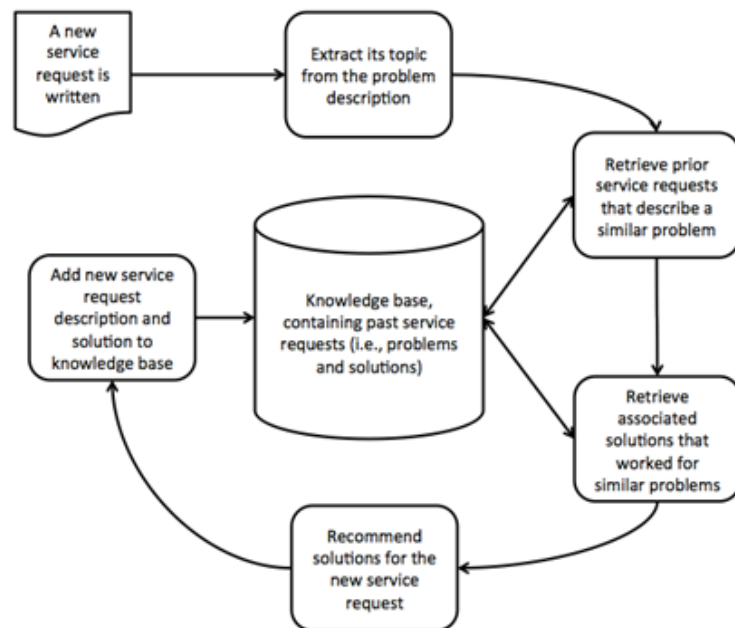
Inventx decided to use text analytics to improve its IT service management process with two goals in mind. As well as automating the process of categorizing and monitoring the steady stream of service requests flowing into its helpdesk, Inventx wanted to embed text analytics directly into the process of solving individual customer issues. Currently, about 60% of incoming service requests are resolved by the helpdesk immediately. However, 40% require input from back-office domain experts. Identifying the appropriate expert and routing requests in a reliable and timely fashion is a non-trivial task, given the complexity of the IT systems Inventx is operating and the importance of these systems for its banking clients. Inventx has implemented a “recommender” system that cannot only suggest the most appropriate expert for solving a problem, but also possible solutions. This system is based on a set of more than 100,000 historical service tickets that include data about who solved which problem, how the problem was solved and how much time the resolution required. Using case-based reasoning, the system derives its intelligence from mining the textual descriptions of these past problems and solutions.

When a new service request arrives, the system makes recommendations based on finding similar prior requests and then identifying experts and solutions that have proven to be efficient and effective in the past (see Figure 4). To achieve this, the outputs of the topic modeling analysis (i.e., the statistical assignment of topics to requests) are used to match new and past service requests—an approach that has been shown to work well for electronic health records.¹¹

However, Inventx learned early on (as did ITS and Hilti) that the inadequate quality of data in its historical service requests hinders the implementation of this approach. While analyzing historical requests, Inventx became aware that the solution description for individual requests was often too short and sometimes even meaningless. Many of the text fields simply stated “task completed,” without specifying exactly how the task had been completed. As a consequence, Inventx started promoting usage of the solution field among service agents with the aim of increasing the quality of its knowledge base.

11 Miotto, R. and Weng, C. “Case-based reasoning using electronic health records efficiently identifies eligible patients for clinical trials,” *Journal of the American Medical Informatics Association* (22:e1), 2015, pp. e141-e150.

Figure 4: Architecture of a Topic-based Solution Recommender



Comparing Inventx's use of text analytics with the other two cases reveals one major difference. Both Hilti and ITS focused on extracting novel insights from their digital data streams, whereas Inventx went beyond that. Based on real-time analysis of its textual data streams, Inventx implemented a set of automated processes, such as the automated routing of requests to the most suitable agent or the data-driven recommendation engine to aid agents in providing the best answers possible.

Interestingly, Inventx was not interested in visualization. Even though its text analytics system is based on the same tool used by ITS and Hilti, the tool is hidden in the back end and not used by human analysts. This approach of using digital data streams to drive, and even automate, decision making has been referred to as "process-to-action." In contrast, ITS's and Hilti's approach can be classified as "assimilate-to-analyze," where new and unprecedented insights are gained into the inner workings and the environment of the organization.¹²

Lessons Learned

The goal of all three organizations that we accompanied on their text analytics journeys was to extract insights from the contents of unstructured digital data streams to better understand their customers' needs and improve their service processes. Despite their differences in the type of service offered, their customers, their industry and their geography, as well as the initiatives that were triggered as part of the text analysis, four general lessons can be drawn from their experiences.

Lesson 1: Position Text Analytics in Business Units, Not IT

Many organizations find it difficult to decide whether their analytics initiatives should be positioned in the central IT department or embedded in a business unit. Prior research has identified four dimensions of data analytics capabilities that can help organizations to make this decision.¹³ The dimensions are dataset, toolset, skillset and mindset, and it has been

argued that an organization's analytics journey typically goes from dataset to toolset to skillset to mindset. This perspective would make the IT department the obvious starting point for running text analytics because it has two of the four capability dimensions: toolset (the products and platforms for analyzing the contents of digital data streams) and skillset (the coding and statistical skills for detecting patterns in data streams). Indeed, until recently, performing text analytics required extensive technical capabilities. Early tools were cumbersome to run, lacked documentation and provided output that lacked intuitive visualizations. This is no surprise, given that early implementations were developed by and for computational linguistics researchers.

Our experience, however, suggests that today's text analytics tools can be used successfully by tech-savvy business people, who typically have a better knowledge about what datasets exist (dataset) and how business value can be generated by harvesting the datasets (mindset). If they have access to sophisticated and user-friendly tools (toolset), either provided in the cloud or in-house, and are able to learn the necessary skills (skillset), we found they are in a better position than IT people to derive findings from text analytics. In fact, across all three organizations we studied, business users had no difficulties understanding and interpreting the outputs of topic modeling algorithms when applied to a dataset they were familiar with. Revealing the first topic modeling results to business users often led to an "aha" effect and triggered lively discussions. They realized that text analytics can enable them to discover in a matter of minutes the thematic structure of more than 100,000 service requests and can provide tremendous opportunities for service improvements.

The advantage of positioning text analytics capabilities in a business unit (e.g., marketing and sales, customer service) is rooted in the fact that analytical projects are less about rolling out IT tools and more about understanding how these tools might be used for creating business value.¹⁴ Business users know best which questions to investigate, which datasets to explore and how

12 Pigni, F., Piccoli, G. and Watson, R., op. cit., 2016.

13 See Barlow, M. "The Culture of Big Data," *O'Reilly Media, Inc.*, 2013; and Pigni, F., Piccoli, G. and Watson, R., op. cit., 2016.

14 For a more detailed discussion of this topic, see Marchand, D. A. and Peppard, J. "Why IT Fumbles Analytics," *Harvard Business Review* (91:1), 2013, pp. 104-112.

to translate insights into actions. Moreover, introducing text analytics into an organization this way generates trust among users and management alike, and ensures that use cases for text analytics are well understood. Of course, most business departments will need help in terms of tools and skills, but the availability of easy-to-use, cloud-based analytics tools makes these challenges manageable.

Lesson 2: Learn the Basics of Text Analytics by Analyzing the Past

Having identified a textual data stream that potentially has high organizational value but has been impossible to analyze with traditional tools, it is time to learn the basics of the text analytics tool. Our research shows that it makes sense to start by analyzing an historical dataset (e.g., last year's service tickets) before trying to gain insights from data streams in real time. This approach provides a way to experiment with the functionalities of the tool and to learn how to read and interpret its outputs. By first analyzing the past, users also learn how to detect patterns and trends in the data—for example, by identifying the most prevalent topics in a body of text and tracking their evolution over time.

To facilitate learning, it is essential to provide opportunities for sharing individual insights derived from the tool's outputs. Whenever text-mining results were presented at meetings in our case organizations, typically with representatives from all levels and business functions present, lively discussions ensued immediately. Everybody had a thought to share; they would either throw in explanations as to why topics spiked or ideas on how to remedy some of the issues identified. Overall, these initial meetings could be described as inspired and animated. Even if the discovered patterns and trends were not necessarily surprising, having empirical evidence—instead of intuitions and anecdotes—was perceived as incredibly useful by all participants and formed a fertile basis for further discussions and ideas. While retrospective analysis might be perceived by some as providing only marginal returns, it is nevertheless an essential stage for organizations to pass through because, without it, they will not be able to fully exploit the text analytics tool.

Moreover, analyzing the past might reveal some unpleasant IT systemic truths. In all three

case organizations, the initial data-extraction and cleaning process revealed what the text analysis later would only confirm: the level of data quality was sometimes inadequate. As described above, the written solutions to service problems were often inadequate, with just “task completed” rather than a detailed description. While educating and training users might go some way toward solving this problem, the existence of inadequate textual data is indicative of something more systemic. For example, two of the three organizations had chosen to implement the best practices model embedded in their respective ERP system and had not considered customizing any major component when they did this. The consequences of this choice only became apparent years later—when the data was heading for text analysis.

Lesson 3: Move Eventually Toward Real-time Monitoring of Textual Data Stream

Once organizations have learned the basics of text analytics and have set up the required technical infrastructure, they typically begin to think about extending the analysis to monitoring textual data streams in real time. Often, they start by re-running text analytics at shorter and shorter intervals (e.g., weeks, then days, then hours) to get a feel for how volatile the data streams are. Typically, in this period, a lot of ideas for exploiting the analytics outputs are generated, and quick wins are often proposed to justify further investments. Some of the organizations we studied developed interactive visualizations for tracking customer issues in real time—for example, when a new product or service was introduced or features of existing products were updated. Others set up alerts triggered when there was a spike in responses for a particular topic.

An essential prerequisite for moving toward real-time monitoring is that the textual data must be automatically streamed into the analytics tool. This often involves implementing adapters to periodically extract data and accompanying metadata from transactional systems, implementing streaming APIs to enable real-time analytics, transforming data structures and loading data into the text analytics tool. All three case organizations found that locating

the appropriate data, querying the appropriate tables and fields, and ensuring that the data set was complete and coherent typically took more time than expected. They had to identify programmers with the skills to access the back-end transactional systems.

Lesson 4: Embed Textual Analytics into Operational Business Processes

Once organizations have the capability to automatically analyze the written word, it is time to translate the insights derived from text analytics into actions. Increasingly, this means making existing business processes smarter by embedding text analytics into the control flow. For instance, algorithmic analysis of the content of service requests can enable the automated routing of tickets to the most appropriate service agent or the recommendation of solutions based on how requests were handled in the past. The objectives an organization decides to pursue at this point in its text analytics journey typically have long-term impacts and are aimed at delivering high business value: changing the nature of the service process.

This stage of the journey entails more than “just” extracting insights from data streams meant to be consumed by a human decision maker—it means changing the mindset of the organization to embrace algorithmic decision making and the automated initiation of actions. In the projects we accompanied, a somewhat philosophical dilemma quickly became apparent. Business managers wondered: *“To what extent are we willing to let technology determine the way we perform our service operations?”* At first, this question was meant to delineate between ideas that are doable in the short term, like building a dashboard, and those that might be conceivable in the long term, such as developing a solution recommender system. After a while, however, organizations understood that topic modeling is just the tip of the iceberg. Combining these algorithms with technologies for speech recognition, machine-learning algorithms for making predictions, and avatars used as virtual service agents has the potential to completely automate the management of customer service operations.

Thus, text analytics holds the promise of redefining business processes just as humans

do as part of a reengineering effort. But unlike humans, algorithms are able to reengineer in real time. Using algorithms to redefine business processes has been described as “Analytics 3.0” or “machine-reengineered processes.”¹⁵ But this approach poses a dilemma for organizations and seems to be an ideological step that organizations—and society in general—are struggling with. After all, what happens if a machine-recommended solution to a critical customer problem ends in disaster? Finding the right middle ground will therefore present a major challenge in the future.

Concluding Comments

Text analytics is a journey. As organizations learn the basics of the tools, they also learn about themselves. Each journey starts by taking a detailed look at historical textual data, continues with monitoring data as it comes in and ends with decisions that change the nature of how work is carried out. We studied three organizations in detail that applied text analytics to the contents of their service requests. While each organization had a different analytical goal, their experiences were surprisingly similar. Each set out to achieve service improvements for its customers, although other business units may equally benefit from text analytics. Today’s advanced text analytics techniques are domainfree—that is, they try to extract knowledge from data, regardless of its content.

About the Authors

Oliver Müller

Oliver Müller (oliver.mueller@itu.dk) is an associate professor in the Information Management section at the IT University of Copenhagen. As part of his research, he studies how organizations can create value with (big) data and analytics. He particularly focuses on the application of methods and tools for extracting knowledge from unstructured data, from both

¹⁵ See Davenport, T. H. “Analytics 3.0,” *Harvard Business Review*, 2013, available at <https://hbr.org/2013/12/analytics-30>; and Wilson, H. J., Alter, A. and Shukla, P. “Companies Are Reimagining Business Processes with Algorithms,” *Harvard Business Review*, February 8, 2016, available at <https://hbr.org/2016/02/companies-are-reimagining-business-processes-with-algorithms>.

social media and enterprise-internal data sources. His research has been published in the *Journal of the Association of Information Systems*, *European Journal of Information Systems*, *Communication of the Association of Information Systems* and various others.

Driving Innovation in a Digital World. He has held various editorial roles and leadership positions in information systems research and education.

Iris Junglas

Iris Junglas (ijunglas@fsu.edu) is an associate professor at Florida State University. Her research interests cover a broad spectrum of technology innovations, but most prominent are the areas of u-commerce (ubiquitous commerce), the consumerization of IT and business analytics. Her research has been published in the *European Journal of Information Systems*, *Information Systems Journal*, *Journal of the Association of Information Systems*, *MIS Quarterly*, *Journal of Strategic Information Systems* and various others. She is a senior associate editor for the *European Journal of Information Systems* and an editorial board member of the *Journal of Strategic Information Systems* and *MIS Quarterly Executive*.

Stefan Debortoli

Stefan Debortoli (stefan.debortoli@uni.li) is an associate researcher at the University of Liechtenstein; he is also the CEO and co-founder of MineMyText.com. His studies focus on big data analytics as a new inquiry strategy in information systems research. His work has been published in the *European Journal of Information Systems*, *Communications of the Association of Information Systems* and *Business & Information Systems Engineering*.

Jan vom Brocke

Jan vom Brocke (jan.vom.brocke@uni.li) is the Hilti Endowed Chair of Business Process Management, Director of the Institute of Information Systems and Vice-President at the University of Liechtenstein. His research focuses on digital innovation and transformation, business process management and data analytics, and has been published in, among others, the *European Journal of Information Systems*, *Journal of Management Information Systems* and *MIS Quarterly*. He has published seminal books, including the *International Handbook on Business Process Management* as well as *BPM -*