BUTLER ANALYTICS

Text Analytics

A business guide

February 2014

Contents

3 The Business Value of Text Analytics

4 What is Text Analytics?

6 Text Analytics Methods

8 Unstructured Meets Structured Data

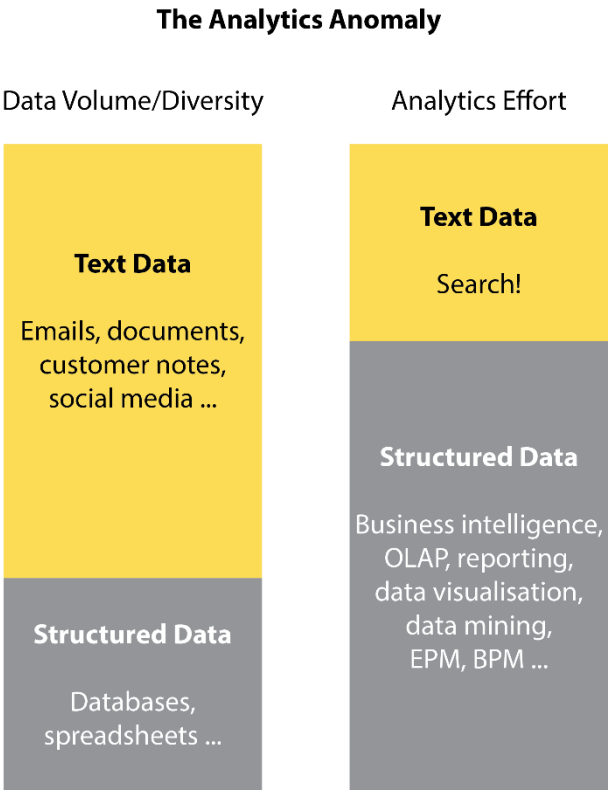
9 Business Application

10 Strategy

The production of this document has been sponsored by FICO Make every decision count.™

The Business Value of Text Analytics

The business value of text analytics is fairly straightforward. The large amounts of text based data that most organizations possess, acquired and managed at considerable cost, can be analysed in such a way that insights can be gained to improve the efficacy and efficiency of business operations. Text based data are an untapped resource in many organizations. Structured data (customer details held in a database for example) on the other hand are very well exploited, primarily because they are specifically formatted for computer processing. While unstructured data, primarily text, is well suited for human communication, there have been significant hurdles to overcome to make it amenable to processing by computer systems. These barriers have been slowly eroded to the extent that significant value can now be extracted from text.



This is something of an irony since text based data typically accounts for eighty per cent of the data most organizations generate and process. Emails, documents, social data, comments and a wide variety of other text are usually archived, but typically not analysed in any meaningful way. The cost of creating, capturing and managing this information is considerable. In a service based business most employees can easily be categorized as information workers, and the cost of the information they generate is directly related to associated labour costs. Viewed in this way the cost of text data in many organizations is in excess of fifty per cent of all costs. Clearly any technology capable of automating the extraction of useful information from these data should be of interest.

The application of text analytics technologies has grown rapidly with increased regulation, the proliferation of social data, and efforts to record the thoughts and comments of customers. Embedded in the terabytes of unstructured data are patterns which can serve a diverse range of purposes, from flagging when a customer is likely to close their account, through to fraud detection. The value of text analytics is amplified when both structured and text data are combined, and to this end text mining technologies are witnessing significant uptake. In this scenario text data are converted into a form where they can be merged with structured data from transactional systems and are then scrutinized by data mining technologies, whose sole purpose is to uncover hidden structure in data and reveal exploitable patterns. It is then crucial that these patterns can be deployed in a production environment, with full monitoring of performance as scoring is performed on new incoming data. Managers will not be confident unless they can assess the benefits a predictive model is bringing on a real-time, ongoing basis.

To realize value from the very large sums invested in creating text data an organization needs to carefully plan and execute a business led initiative. This involves identification of business processes

where text analytics might add value, the creation of text analytics capability, and a feedback loop in which information capture is informed by the outcome of analytics processes. This latter point is crucial, but somewhat surprisingly is often not mentioned by suppliers and consultants in this domain. If a certain type of information generates useful patterns then it becomes important to understand why, and attempt the capture of other information which might amplify the value of the analytics process.

Underlying all of this is some fairly simple economics - the cost of discovering and exploiting information derived from text analytics should be less than the value realized. Fortunately analytics often produces measurable outcomes captured by metrics such as lift. A two per cent increase in lift can mean a very considerable return on text analytics investments in many customer and marketing oriented activities.

Finally it should be noted that the scale and scope of text analytics will be accelerated by the current developments in big data technologies. The most heavily visited topic on the butleranalytics.com web site is text mining. We predict that this will become the largest growth area within the data analytics space, and a key differentiator in the benefits organizations reap from their analytics activities.

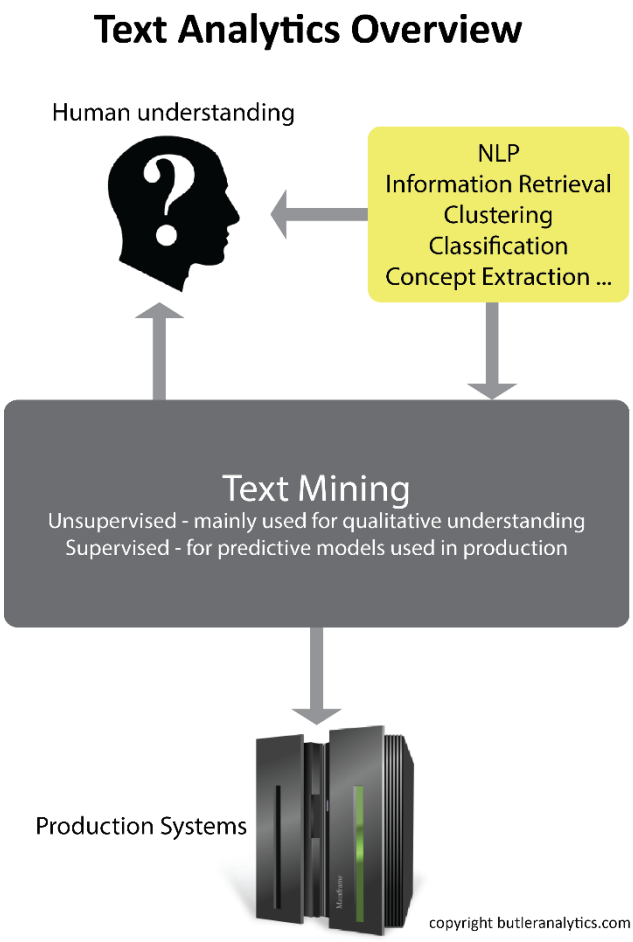
What is Text Analytics?

Text analytics convert unstructured text into useful information. This information can be in a format suitable for human consumption (categorized documents for example) or fed into computer systems to improve business processes (detecting customers who might defect). There are many techniques used in text analytics, but the target for the resulting information is always a computer system or people who need the information.

The information that text analytics can deliver to a person is very diverse. This ranges from language translation through to identifying important entities (people, places, products), categorizing documents, identifying important topics, establishing links between entities, establishing document similarities and so on. Much of this functionality comes under such headings as natural language processing (NLP), information retrieval, information extraction and several other domains which are still strongly associated with their academic roots. As far as the user is concerned this form of text analytics should simply reduce the overheads associated with finding and processing information, and many commercial products exist that perform exactly this function. Various surveys show that the average information worker spends up to a third of their time locating information and trying to make sense of it. If text analytics can reduce this overhead by just a few per cent, then simple math would show that the savings are considerable. In reality text analytics delivers much more than just a few per cent improvement, and tens of per cent improvement is common.

Processing unstructured text data so it can be processed by computer systems is a wholly different exercise. Powerful data mining algorithms, capable of identifying patterns within data, do not understand unstructured data. To this end many of the techniques mentioned above (NLP, concept extraction …) can be used to extract features from text (important entities for example) which can be used as input for the data mining algorithms. These features are often combined with structured data to provide additional insights into behaviour. Text data in the form of customer notes might be processed to deliver features that show important terms used by customers, and when combined with customer records from a database will often improve the accuracy of patterns found. These might indicate suitable targets for a marketing campaign, or potential delinquency. The terms used

for this type of activity are ambiguous, but for our purposes we can call this text mining and seen as an extension of data mining.



While text mining is often used to identify patterns which can be used in production systems, it too can provide output suitable for human consumption. This type of mining is called unsupervised learning - the data are simply fed into the algorithms and the output shows various clusters of documents, possibly revealing significant insights. A second type of text mining is more concerned with finding patterns that improve business processes through deployment in computer systems. This is called supervised learning where the text mining algorithms learn from a large sample of text data, and the resulting patterns are usually tested against new data the resulting pattern hasn't seen before. These patterns often classify new data (risk or no-risk for example), create probabilities of new data being in a particular class, or calculate a numerical value for new data (a credit limit for example).

In summary text mining offers the potential to automate the analysis of text data and feed resulting patterns directly into production systems. Many other techniques exist to process language for human consumption, although some of these techniques can also provide input to business processes. Text mining employs many machine learning technologies, and since this is a domain of intense interest, it is here that many advances will be made. Coupled with the advances being made in the storage of text data (column databases for example), the use of text mining technologies will see accelerating uptake over the next few years. Of course the adoption of such technologies can happen through in-house initiatives or by employing ready-made solutions. As always the best route for many organisations will be the middle-way – technologies that address much of the problem at hand, but with a sufficiently powerful toolset that bespoke work is not problematical.

Text Analytics Methods

Natural language text is not a medium readily understood by computer systems, in contrast to the neatly arrange rows and columns in a database. This is the primary reason that text analytics has had such a long gestation before it could be usefully employed in a business arena. It also means that much of the effort involved in text analytics is preparatory work, to make sure the data are in a format that can be processed by text applications.

The first stage in dealing with text data is nearly always the process of identifying individual words and phrases (usually called tokens). Even this is not as simple as it sounds since abbreviations, acronyms, synonyms and ambiguity make the task quite involved (the word 'wave' has multiple meanings for example). It is also usually necessary to identify 'parts-of-speech', and specifically which words are nouns, verbs, adjectives and so on. Many words are meaningless as far as text analysis is concerned and can be 'stopped out'. Words such as 'a', 'it', 'and', 'to' and so on can usually be stopped and unsurprisingly are called stop words. A significant part of natural language processing is dedicated to these tasks, and it is a prerequisite before other work can be done. At the heart of this approach is an attempt to infer some level of meaning within documents (identify important entities, concepts and categories).

A wholly different approach can be adopted by using statistical methods. Here we might simply count the number of times various words appear in a corpus of documents and infer some level of importance from the resulting frequencies. One of the most useful metrics based on this approach is called the inverse document frequency. This increases in importance as a particular word appears frequently in a given document, but is not common in all documents. The word 'loan' may appear frequently in a corpus of documents and have no particular importance in a particular document. Whereas the word 'default' would appear less often (hopefully) and have more significance in a specific document. This approach can give useful results, but context and meaning is almost entirely lost. In an attempt to address this, short sequences of words called n-grams can be processed. This does at least offer the opportunity for frequent combinations of words to be identified. Significantly more sophisticated approaches are often used in commercial text analytics products, a good example being probabilistic latent sematic analysis where documents can be assigned to discovered topics.

Mary had a little lamb
Its fleece was white as snow
And everywhere that Mary went
That lamb was sure to go

Stopped words - a, its, as, and, to, go

Important nouns - Mary, lamb
Frequencies - Mary (2), lamb(2), fleece(1) ...

Context - Mary and lamb appear twice within 5 words of each other.

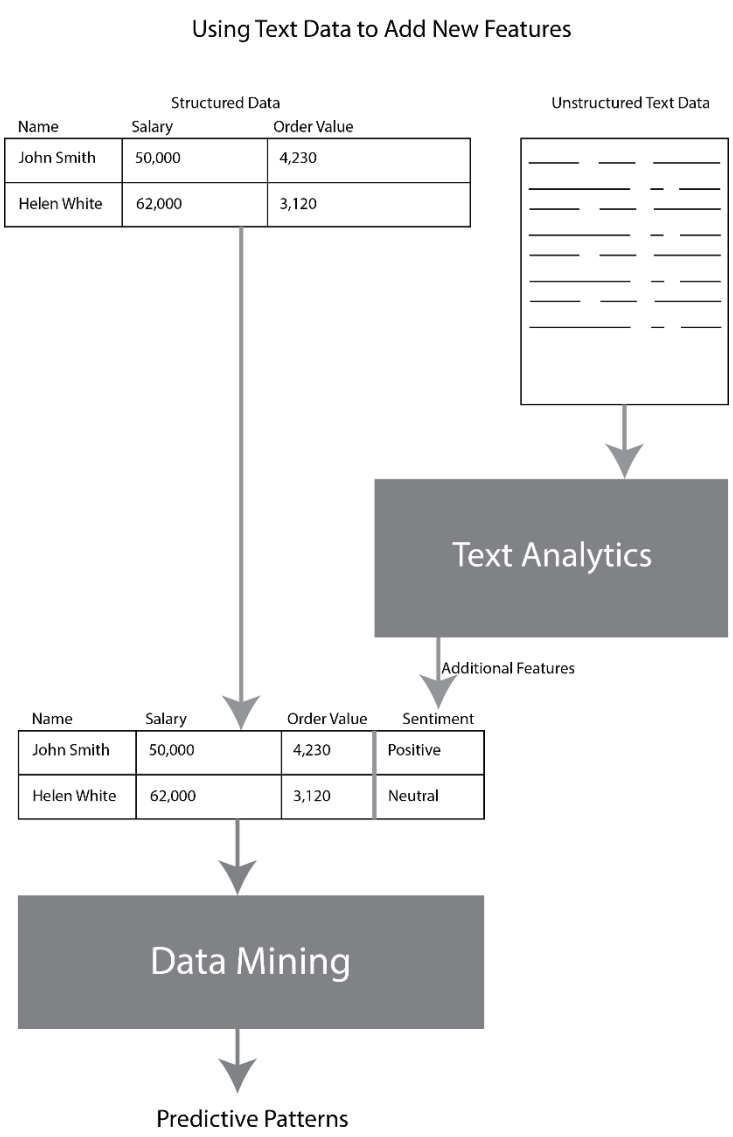
The above methods, and many others, can be used to generate input to data mining activities. We might have a detailed transactional history of customer activity, but with little notion of how happy or otherwise the customers are. To this end we might use some of the above methods to identify customer sentiment and add additional variables (usually called features) to our customer records. This approach is proving to be successful in many applications.

There are two ways to address the complexities associated with text analytics. The first is simply to buy a 'solution' for the task at hand. Various suppliers provide analytics solutions to a range of vertical and horizontal business needs. The benefits associated with this approach include fast time to implementation, reduced need for in-house staff and the ability to call upon a supplier that has experience in a particular domain. The downside is usually lack of tailoring to particular needs, less awareness of how an application actually works, and a potential dead-end if a solution cannot be modified sufficiently. The alternative is to build an in-house facility with associated costs, but with the opportunity to address specific needs accurately.

Text analytics can deliver simple to use functionality, often seen in information retrieval and characterised by some form of search capability. But it is making its way into data mining activities and it is here that more capable organisations will realise significant advantage.

Unstructured Meets Structured Data

Text data are typically held as notes, documents and various forms of electronic correspondence (emails for example). Structured data on the other hand are usually contained in databases with fixed structures. Many data mining techniques have been developed to extract useful patterns from structured data and this process is often enhanced by the addition of variables (called features) which add new 'dimensions', providing information that is not implicitly contained in existing features. The appropriate processing of text data can allow such new features to be added, improving the effectiveness of predictive models or providing new insights.



Incorporating features derived from customer notes, email exchanges and comments can improve lead targeting, flag possible defection and even contribute to the identification of fraud. The methods used to extract useful features from text depend on the domain, the nature of the business application and the characteristics of the text data. However a statistical approach based on a frequency matrix (a count of words appearing in various text sources) often yields useful new features after the application of appropriate statistical techniques. Other techniques might employ named entity extraction (NEE) where probabilities can be assigned to the likelihood that a particular document refers to a given entity (people, places, products, dates etc.).

A prerequisite for combining text and structured data is some form of integrated data environment, since the sources of a data could be highly diverse and volatile. While building predictive models can be facilitated using various interfaces to unstructured data, implementing the resulting models requires tight data integration and a scalable computing environment. This can be achieved through big data infrastructure such as that offered by Hadoop and associated

technologies, although this is definitely not a trivial undertaking. The alternative is to embrace integrated infrastructure and tools provided by some of the larger suppliers in this domain.

Such is the complexity of integrating text analytics with structured data that most organisations will opt to buy solutions for their needs. There is no point reinventing the wheel here, and some advanced solutions are already emerging for customer and marketing applications where text data are incorporated into data mining activities. Probabilistic latent semantic analysis and specifically Latent Dirichlet Allocation is a highly sophisticated technique used to associate documents with a topic. Specialist knowledge is needed to employ such techniques and many businesses will simply opt to buy capability rather than explore such highly complex statistical methods. The techniques used are just a small part of the story with infrastructure, skill sets, presentation methods, management and performance monitoring representing the larger piece of the cake.

The integration of text data sources with structured data will see significant progress over the next few years. Organisations that are willing to integrate the missing 80% of their data that text represents (missing from current analytical activities) will gain insights and operational improvements that would otherwise not have been possible.

Business Application

Text analytics has wide application in a business environment. For users wishing to reap productivity gains, text analytics can automatically filter spam, categorize messages (emails for example), label documents and enable fast and relevant information search. This can be viewed as the 'traditional' role for text analytics, although more contemporary applications include fraud detection, warranty claim processing, competitor analysis and social media monitoring. Beyond this text data are being used to add new features to data mining activities with the aim of creating predictive models which might be used to detect possible customer defection, new selling opportunities and delinquency. It can also be used to provide new insights into customer behaviour through identifying archetypes.

The potential benefits of text analytics are not 'automatic'. Domain experts are needed to provide input to the analytics process and sanitize results. This applies to almost every application of the technology regardless of whether a 'plug and play' type solution is being used or a bespoke application has been built. Even web based services which provide sentiment analysis of social media require considerable amounts of configuration if results are to be meaningful.

The focus for text analytics solutions is primarily in the customer and marketing domains. Such solutions are often cloud based, but for larger organizations a in-house deployment might be necessary because of latency and security issues. Either way text analytics provides insights into customer behaviour that are not accessible through analysis of the structured data contained in databases. This extra dimension can be used to tailor a more relevant interaction with customers and predict future behaviour. For example it may be possible to identify which customers are perfectly good credit risks, but sometimes make later payments because of lifestyle. It is much more likely that a customer 'sweet spot' can be identified through text analysis than by any other mechanism, since text contains significantly more potential information than the history, demographics and other data held in databases.

In a marketing environment, text data in the form of open surveys (where respondents can add free text comment), can be used to extract nuances which simply cannot be accommodated in a closed response form. This might enable sentiment scores to be created or the identification of terms and

concepts that had not been anticipated. Obviously this is closely related to the sentiment analysis of social media, which at the current time is over-hyped, but is quickly evolving to provide behavioural insights and trend analysis for marketing activities.

While customer and marketing applications might be the most obvious ones, text analytics applies to any domain where text data is acquired. In a manufacturing environment the notes made by maintenance staff might be used to improve the prediction of maintenance requirements and the avoidance of down time. In a medical environment text notes that are captured during the diagnostic process can provide valuable input to understanding patient concerns and the process of diagnosis itself.

Perhaps the most promising application of text analytics is the creation of new features for data mining processes. Combining structured and unstructured data in this way facilitates the meeting of two quite different information 'dimensions' and is already being used in sales, marketing and customer care applications.

As always business management will be tasked with the need to identify opportunities and decide how unique they want a solution to be. Packaged solutions potentially reduce risk, but also reduce opportunity. A bespoke solution introduces technical and business risk, but also provides the most opportunity. Fortunately a number of suppliers offer a middle way with many of the technical, architectural and business risks largely mitigated, but with an opportunity to deliver a tight fit with individual business needs.

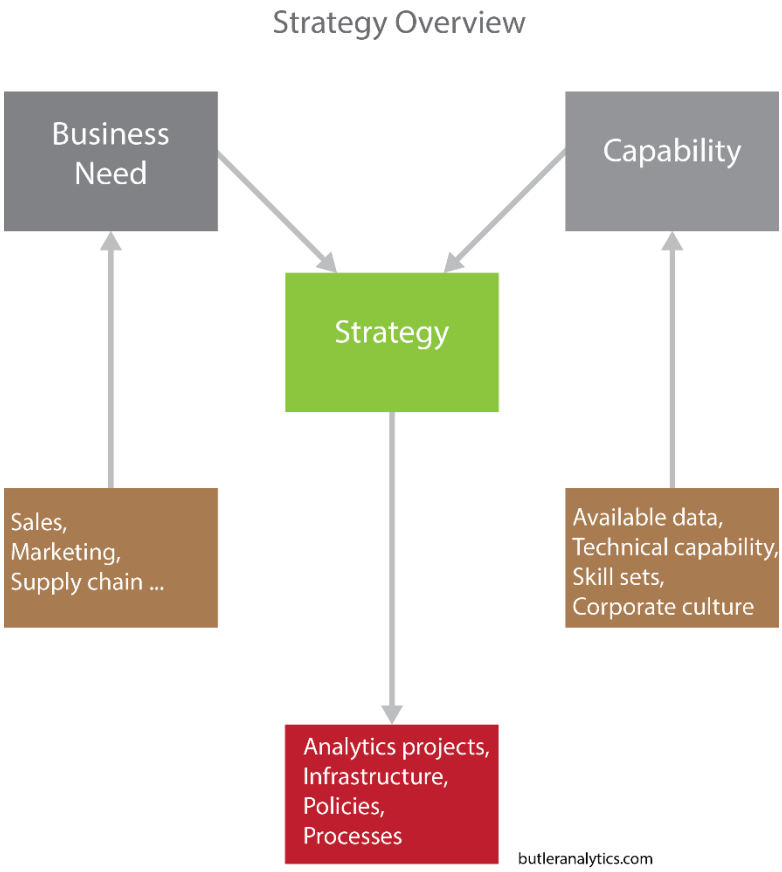
Strategy

Strategy is always the meeting point of need and capability. The starting point in a text analytics strategy is the identification of business processes where text analytics might deliver benefit, and an awareness of what is possible. Business processes that are candidates obviously need access to relevant text data. At the current time this usually applies to customer and marketing applications, although as text analytics becomes more prevalent so businesses will collect more text data to enable the analytics process. As far as capability is concerned it is usually a given that provided good quality text data is available, so value will be created through the analytics process, although this is nearly always an iterative process.

Once candidate business processes have been identified it becomes a matter of fleshing out some form of cost/benefit analysis. With analytics technologies it becomes much more difficult to estimate benefits unless data are available from other organizations, or suppliers who have experience in the domain. Typically, an increase in lift (value created with new information divided by value created without information, multiplied by 100) of just a few percent will more than adequately reward many text and data mining activities. The cost of developing and deploying text analytics applications depends very much on the route taken. Text analytics may be part of a much larger 'big data' project, in which case it becomes more difficult to allocate costs. Costing discrete projects is usually much easier.

This does not mean however that determining costs will be necessarily straightforward. Unlike traditional process automation projects (e.g. ERP, CRM) where deployment is essentially linear, analytics projects are usually iterative in nature. Again it is very useful to have access to people who have knowledge of building and deploying analytics processes, although the nature of text data will be particular to each organization and this inevitably introduces some variability. The cost

components will include hardware and software (unless a cloud service is used), skill sets, domain expert involvement, performance management and monitoring and possibly the acquisition of new data. This latter point is more important than is immediately obvious. Sourcing external data (social data, market data etc.) is an obvious cost, but it is more than likely that the results of analysis will imply that greater data capture when dealing with customers for example might deliver more accurate insights and predictive models. There is a cost associated with this and it needs to be taken into account.



Finally there is a cost associated with the management and monitoring of processes involving analytics. The insights and models derived from analysis usually decay with time, simply because markets and customers change. Monitoring performance and feeding this back into the analytics process is not a trivial matter and will impose its own overhead.

While we have been conservative by suggesting just a few percent increase in lift, it does happen that the benefits can be considerably greater than this. A more useful model for modelling the return from an investment is Expected Return. This allows for multiple scenarios and will generate an expected return from an investment. While this is not widely used at the current time other than in some specific industries (petrochemicals for example), it does give a good feel for risk and is more appropriate for analytics projects where there are more unknowns.

Analytics projects do need a somewhat different approach to risk management than traditional IT systems. It really is not enough to develop a model and leave business users to get on with it, the whole process needs much finer integration between business, IT and data analysts.