Best Practices Guide

Six steps to successful text mining

Text mining helps organizations streamline business processes and overcome challenges by gaining insights from their mountains of unstructured textual data. However, as with any data science project, some essential steps must be followed to produce successful results. This Best Practices Guide offers six tips to help organizations get the most out of their text mining projects.

opentext™

Contents

Introduction: There's more to text mining than "Siri, find the good stuff for me."

These days, many organizations are challenged by a lack of quick access to the right information to improve decision-making. They know it is somewhere within their terabytes of internal and external data, but are unable to find it quickly and cost-effectively. Number crunching alone will not tell them what they need to know. Often, the answers lie in unstructured data, such as written text.

Text can be especially challenging to analyze, as it reflects the ambiguity, richness and variety of human speech. Moreover, business context can vary widely from one organization to another. For example, "hedge" means entirely different things in landscaping, finance or law. With text mining, a subset of artificial intelligence, organizations can navigate through millions of pages of content, a task that has become far more complex than mere indexing or keyword searching.

Modern, AI-driven content analytics solutions, such as OpenText™ Magellan™ Text Mining, have evolved to handle sophisticated processes and evaluate the "aboutness" of text. Ranging from a single sentence to a whole collection of documents, they can understand the text's sentiment and level of subjectivity. They can also evaluate their own accuracy, providing feedback to guide an efficient yet secure workflow process.

Setting up an enterprise text mining project can seem intimidating to first-time users because it is a relatively novel and complex technology. Misunderstanding or misalignments can occur between stakeholders, including executives, IT departments, specialized linguists and data scientists setting up the project, metadata specialists who—before the advent of text mining applications—would have processed the text manually, and Line-of-Business managers running the project and using its output.

This guide offers best practices to simplify the process and minimize misalignment based on OpenText's experience in building industry-leading, AI-driven analytics solutions that leverage Magellan Text Mining, such as Magellan for Intelligent Recommendations and AI-Enhanced Voice of the Customer Analytics, and in helping our customers derive more value and insight from their enterprise content.

Precision

In document discovery projects, refers to finding only relevant documents within a set, with no unwanted documents.

Recall

The converse of precision, refers to finding all relevant documents, not overlooking any (even if a few irrelevant ones are also returned).

Step 1: Diagnose the problem and define the goals

Evidence of business struggles are typically apparent in an organization's bottom line. But research and examination are required to trace the problem back to the cause. For example, losing $1.2 million in a quarter could be due to underpricing products or because the company did not effectively target the right customers. Losing 350 customers in a month might be attributed to a competitor releasing a better-performing product or to frustrated buyers not being able to find needed support.

The first step toward using text mining to solve such problems is to associate business solutions with their challenges. For example, an organization may seek to become more competitive by categorizing customer feedback to gain better insight into what customers want or to improve its indexing team's productivity by implementing a semi-automated document tagging solution.

The project team needs to go even further by setting goals that are quantitative, measurable and reasonably achievable. For the examples above, the team could set a goal to use six categories of customer feedback, based on communications with the customer service department or to improve indexing team productivity by at least 20 percent over the last quarter.

Step 2: Figure out how text mining can help

Organizations should focus on defining goals that are quantitative and business-focused.

For example, aiming to achieve 90 percent recall with the text mining solution is not a business solution benefit, since it does not directly solve a business problem. Some business problems may be solved if precision and recall are at 80 percent, while others could be overlooked even if they reach 95 percent.

Organizations should instead set business-focused goals, such as improving the bottom line, and then establish criteria to help get there.

However, even when organizations know exactly what the business problem is and the changes they want to make, they may not immediately realize how (or even if) text mining can solve it. The proposed solution may not be obvious to organizations unfamiliar with text mining.

The solution? Discuss problems and potential solutions with a trained semantic analyst, for example in the OpenText Semantic Strategy Workshop described at the end of this guide.

Step 3: Communicate goals and set expectations

The reasons for using text mining differ, based on different roles in an organization. Executives and managers are more likely to focus on cost reduction, profit growth and/or productivity improvements. Editorial teams, content managers and metadata experts tend to focus on accuracy.

Organizations should set out to manage expectations at all levels. Desired outcomes should be tied to the business goals and solutions defined in Step 1. To be successful, the project team should define expectations when it begins capturing stakeholder needs. It helps to communicate the value of text mining and how it benefits the organization throughout the project.

Typical misunderstandings and discrepancies in the perceived value of text mining include:

Role	Fear	Reality
Metadata specialists:	Text mining will replace us.	Text mining in situations where specialists have been manually adding metadata are typically deployed to improve productivity and consistency and ensure content is tagged more thoroughly and consistently, not to reduce head count. The benefits of text mining deployments include the reassignment of some specialists to more creative, less repetitive tasks.
Metadata specialists/ content managers:	Text mining will never be as insightful and context-sensitive as us.	No computer system can be as flexible and sophisticated in its understanding as qualified human specialists but this is not the goal of text mining. Instead, tools, such as confidence level thresholds, can be used to pick out the most difficult or context-sensitive cases and direct those to experts for manual review.
Metadata specialists/ editorial teams:	Text mining will never be accurate enough.	Linguistic-based text mining solutions can achieve production-level accuracy through intermediate steps of testing early findings against human review and adjusting the parameters as needed.

Understanding common misconceptions and countering them with the benefits of text mining will help project teams set expectations and gain buy-in, which is critical to the success of any AI project. It is important to distill concerns early through training and education and make data specialists part of the solution early in the process.

Step 4: Make detailed plans

To ensure success, include relevant teams and individuals at the following milestones:

Integration plan

This stage should include metadata specialists, content managers and IT managers/ specialist teams to answers questions about the project, including:

- Business goals.
- The key audience.
- Technologies used.
- New workflows.
- Which content should be involved and how it should be prepared to ensure accuracy.
- Where/how the metadata will be stored.
- Anticipated impact on existing products.
- What controlled vocabularies should be trained/automated.
- How the text mining solution should be configured.

Deployment plan

This stage should include executives, managers, metadata specialists, content managers and IT managers/specialist teams to answer questions about the project, including:

- Details of the web user experience.
- Details of the back-office user experience.
- New syndication/content billing procedures.
- New products to be created.
- How to train users on the new workflow.

Step 5: Establish and maintain technical parameters for the text mining project

Text mining is not a one-time operation. It improves over time as the context of the content grows or changes or as the organization of that content evolves. Moreover, organizations often use text mining to support ongoing processes, such as searching internal resources or routing forms to the appropriate destinations. It is important for data scientists and content experts to carefully define the key technical parameters of the project, then maintain them during its lifespan.

Strategic maintenance parameters to focus on include:

- Review of controlled vocabularies (e.g. taxonomies, authority files).
- The impact of new content sources, types and products.
- The impact of new syndication targets.
- The impact of the deployment of new internal/external applications.

Note that Magellan Text Mining can not only extract semantic metadata, it has the knowledge of its own accuracy, so users can depend on its self-evaluation mechanisms to inform an efficient yet secure workflow process. Confidence and relevance scores are useful in a business process workflow, as they can help route the automated or semi-automated process. For example, if the text mining system is 80 percent confident that it tagged the document correctly, a user might want to automatically push that along to the next step in the process. When the system marks something with 60 to 80 percent confidence, the user might direct it through another step where reviewers validate the tag. Using these scores offers organizations the flexibility to apply thresholds that reflect their specific precision and recall tolerance levels.

Step 6: Measure output against business metrics

Performance measurement lies at the heart of any improvement. The impact of deploying the text mining solution should be measured, at agreed upon intervals, against established business problems identified earlier. Depending on the industry and kind of problem an organization is trying to solve, key business metrics may include:

- Customer satisfaction.
- Customer retention.
- Product/service quality.
- Market growth rate.
- Revenue and/or profits.
- Process quality and capability.
- Productivity (including speed, capacity, number of users/customers/served and more).
- Organizational, infrastructure and stakeholder capability improvements.

opentext™

Conclusion

Creating a fruitful text mining project involves more than the actual content analytic algorithms. It requires thinking about the business problem, communicating goals, involving the appropriate stakeholders and finding the right ways to measure success. To learn more about Magellan Text Mining, click here.

To further explore how text mining can unlock value, consider OpenText's Semantic Strategy Workshop. Participants work with an OpenText computational linguist on site to get an overview of how Magellan works and explore various content challenges that it can address. For information, email PortfolioAnalyticsPS@opentext.com.

Learn more from the Magellan Text Mining demo.

About OpenText

OpenText, The Information Company, enables organizations to gain insight through market leading information management solutions, on-premises or in the cloud. For more information about OpenText (NASDAQ: OTEX, TSX: OTEX) visit: opentext.com.

Connect with us:

- OpenText CEO Mark Barrenechea's blog
- Twitter | LinkedIn

opentext.com/contact