Carrier Management

Decoding the Hidden Value of Unstructured Text Data

Executive Summary: **Sentiment analysis and latent semantic indexing are two of the text mining techniques that can help claims handlers unlock the hidden value of unstructured text data, improving prediction accuracy and creating decision-making engines that more closely match human performance.**

By Jason Rodriguez

Data science is about supporting decisions with insights, advice or predictions based on rigorous analysis of prior business facts, decisions and outcomes. One of the biggest challenges for insurers in becoming more data-driven is getting access to a sufficient volume of reliable data in a usable form (Figure 1).

The most usable form of data is a table where each row represents an observation (i.e., a policy, a vehicle, a submission, a claim) and each column contains understandable and useful attributes of those observations or an outcome. Data in this form is commonly referred to as structured data. Unstructured data refers to all other formats and includes text documents, images and sensor data.

It takes foresight and discipline to structure information at the time it is collected from text, images and sensors in order to apply advanced analytics. As a result, structured information captured in IT systems only represents a portion of the information weighed by decision-makers. This information gap between the data available for analysis and the data available to decision-makers presents a fundamental barrier preventing insurers from using machine-learning algorithms or artificial intelligence (AI) to fully support human decision-making.

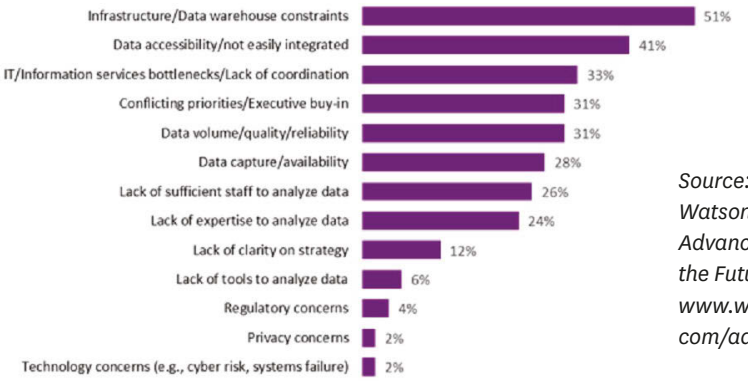
The good news: Most of this missing information can be harnessed from unstructured data sources. Text streams from reports, emails and internal memos offer a treasure trove of potential for data scientists.

Why Is There So Much Unstructured Information?

It turns out most information in the world is unstructured. We try to use technology to standardize communication as much as possible—especially in highly repeatable and fast-paced work, like claims handling—but everything that does not have a place in a standard form with fixed fields will need to be captured somehow. For most other work, the basic facts are presented in long-form reports and other unstructured data sources, which makes text streams one of the most valuable resources available to insurers.

Jason Rodriguez is Data Science Lead in the Insurance Consulting Technology practice (ICT Americas) of Willis Towers Watson, based in Philadelphia. Reach him at jason.rodriguez@willistowerswatson.com.

Figure 1. Insurers' biggest challenges to becoming more data-driven



Source: Willis Towers Watson 2017/2018 Advanced Analytics and the Future survey, www.willistowerswatson.com/aasurvey2018

Text mining techniques can be used to provide passive support to readers by highlighting specific keywords, providing summaries, extracting concepts and even detecting emotional content.

Limitations of Structured Fields on Surveys and Forms

There is a natural inclination to collect free text information, just because we can and because it is easy to store. Think about a survey. There are usually some questions with multiple choice or yes/no answers (which represent structured fields once they are recorded in a database or spreadsheet). But you will nearly always see fill-in-the-blank or even completely freeform note fields.

Those free-text fields represent the survey author acknowledging they want to capture possible answers they did not consider or even ones related to questions they did not ask. Those can be the most valuable fields in the survey because they can contain information we either did not know or even think to ask. Information of this type is captured all over the insurance value chain in emails, phone calls, reports and notes.

The Opportunity in Free Text

Data sources are valuable when they are detailed, varied, tied to business outcomes and relatively unbiased. Free-text information captured in claims often have all of these properties. The claims handling process (especially for long-tail lines) often involves reports and input from a variety of parties. These raw data inputs are usually stored for future auditing purposes.

Claims adjusters are required to peruse these documents efficiently and to consistently make the best decisions for a claim and set the correct case reserve. The larger the volume of information, the more difficult it is for claims adjusters to handle and act on it in a consistent manner, which results in missed information in the claims handling process. Insurers are aware that key information can be stored in these documents, which is why it is one of the top data sources they hope to tap into over the next two years (Figure 2).

Figure 2. Growth in anticipated usage of data sources by personal and commercial lines insurers

Personal lines	Now	Two years
Smart home/smart building data	0%	52%
Usage-based insurance/telematics	26%	70%
Social media	26%	52%
Unstructured internal claim information	39%	61%
Unstructured internal underwriting information	30%	52%
Images	13%	35%

Commercial lines	Now	Two years
Unstructured internal claim information	46%	92%
Other unstructured customer information	11%	54%
Unstructured internal underwriting information	25%	39%
Usage-based insurance/telematics	11%	47%
Web/clickstream/phone/email customer interactions	11%	36%
Images	3%	39%

Source: Willis Towers Watson 2017/2018 Advanced Analytics and the Future survey, www.willistowerswatson.com/aasurvey2018

Text Mining Illustrative Example

The clearest insurance application in text mining is in the claims department. Each claim—especially those that involve bodily injury—can be complex, resulting in a large volume of reports and emails. A claims adjuster's ability to read, understand and quickly make decisions based on this information will be tested, and it is not unreasonable to expect that even the best adjusters will make mistakes.

Natural language processing (NLP) can help claims adjusters manage their time by handling unstructured information in a number of ways. (See related sidebar, "Text Analytics and Text Mining Technologies.") These techniques can be used to

continued on next page

provide passive support to readers by highlighting specific keywords, providing summaries, extracting concepts and even detecting emotional content.

Overcoming Cognitive Biases

Behavioral economists have demonstrated that time pressure tends to exacerbate human perceptual errors caused by cognitive biases. These biases represent departures from logical thinking that the best decision-makers usually find ways to repress.

The famous economists Tversky and Kahneman realized the best way to combat human biases was to use historical information as a reference point (for example, Tversky and Kahneman, 1974, "Judgement under Uncertainty: Heuristics and Biases," *Science,* Vol. 185, pp 1124-1131). This is the exact goal of predictive modeling; the process of building a predictive model involves formally capturing the patterns in prior experience and using them to predict how new observations will materialize.

Claims adjusters naturally take this approach as well. When evaluating a new claim, they may use similar claims from their personal experience as the basis for their decisions. The difference is that a predictive model is transparent (i.e., you can figure out why a decision was made even years later), makes the same decision every time and reaches its conclusions almost instantaneously.

NLP techniques provide useful tools for text mining, but they will not read text as well as the claims adjuster. Natural language *understanding* is an exclusively human trait at this point. Instead, text mining approaches are largely designed to extract certain types of information from text. For example, sentiment analysis involves measuring the sentiment or valence in text. (Editor's Note: Valence refers to the polarity of a piece of text—whether it is positive, negative or neutral.) One form of sentiment analysis simply assigns a numerical value to a body of text depending on the *negativity* or *positivity* of the content. This is useful when trying to estimate the author's emotional state.

Data sources are valuable when they are detailed, varied, tied to business outcomes and relatively unbiased.

Other text mining techniques can help summarize content by using the words contained within the text. Our claim handling example notes that a company produces tens of thousands of unique terms over 10 years, so algorithms like latent semantic indexing (LSI) attempt to reduce this complexity with an approach similar to grouping correlated terms together. The idea is that correlated words could potentially represent similar meaning or concepts, so characterizing a document or a claim by the prevalence of a group of words may be a convenient way of numerically representing each document.

Topic modeling likewise looks for correlations in words to identify themes in the documents. A big advantage of topic modeling is that words can be present in more than one theme (or topic), an idea consistent with natural language.

Each of these text mining approaches results in a set of numbers that characterizes a document (i.e., sentiment score, LSI value or topic proportions), which can easily be incorporated into downstream predictive models. This information can often lead to improved prediction accuracy and, most importantly, decision-making engines that match closer to human performance.

Unlocking the Potential of Text

Unstructured natural language contains much of the historical record an insurer desires to use to improve decision-making with high-quality predictive modeling algorithms. NLP tools available from computational linguistics researchers offer insurers access to this trove of information. The benefits insurers find will be based on the quality of information they already have captured in their structured data and, of course, their text notes. Claims is often the best opportunity for insurers to unlock value. CM

Text Analytics and Text Mining Technologies

Text analytics and text mining technologies are a set of statistical, linguistic and machine learning tools that can help extract essential information that can be missed from unstructured text sources. Natural language processing is a subfield of computational linguistics that is particularly useful for text mining applications.

Natural Language Processing (NLP)

All unstructured data referred to in this article is produced by people and called natural language. A natural language is just a human language, which has a structure, but this structure is not easily understood by computers. Computational linguistics are used to define a mathematical structure within natural language text to enable computer programs to reliably retrieve information from unstructured text. This is a complicated task because human language is highly variable and multifaceted.

While no insurer should expect to push the envelope in the field of computation linguistics, insurers can benefit from the application of standard and emerging NLP techniques in order to extract useful information from their own natural language data.

Text Mining Technologies

Technologies that employ NLP techniques require special consideration.

Computational linguists often publish their work in open-source software like the Stanford CoreNLP library. Technology companies have also published their own open-source text analytics libraries, for example, Facebook's fastText. These libraries represent a specific set of algorithms for dealing with text, which do not overlap with the common libraries for machine learning and AI. CM