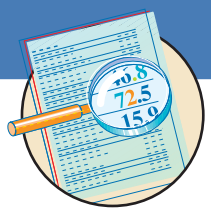

KEY POINTS

- Sentiment analysis combined with complex event processing can provide competitive advantage in forecasting.
- Volume, velocity, variety, and veracity are key considerations when automatically feeding unstructured data into trading algorithms.
- Advances in data analytics soon will enable machines to read and interpret in a way similar to human ability.

Next-Generation Analytics

Can unstructured data improve forecasting?

BY SHERREE DECOVNY

Investment professionals trying to understand trends in a particular stock's behavior typically rely on technical and fundamental data for that stock. But some are trying to improve forecasting by integrating unstructured data including news items, blogs, and Twitter feeds. While the technology is still in its infancy, many believe it is the wave of the future.

Asset managers recognize the benefits of balancing portfolios on an intra-month basis instead of monthly or quarterly. Moreover, they are finding the need to use more real-time data and indicators to tell them when to move on a strategy. A manager might have a long position on a company based on growth expectations, and the first signal that growth is happening might come in the form of unstructured data.

To this end, some firms are starting to use customized news and sentiment analysis, which indicates whether a story is positive or negative. Others are going one step farther and abstracting quantitative structured data from unstructured sources.

"If you wait for an analyst report to tell you that the company is growing or if you wait for a financials call, obviously, you're going to be late," says Richard Tibbetts, chief technology officer at SmartStream Technologies, a platform for real-time data processing.

In the simplest framework, a dictionary of words is classified into various categories such as positive or negative sentiment, fear, greed, joy, and happiness. The occurrences of each type of word or phrase are counted and assigned a numerical score. On a more sophisticated level, natural-language processing techniques are applied so users do not get fooled by phraseology such as "this was *not* good news" or "the results were *not* as strong as we would have liked."

"Being able to take into consideration context and negation and things like that is the next level of complexity that not everybody incorporates in their unstructured text processing but can be very valuable," says George Bonne, director of quantitative research at Thomson Reuters.

The impact of a comment varies according to who says it, so social network analysis can determine whether the comment is coming from leaders or followers, influencers or non-influencers. Analytics software can take blogs, Twitter feeds, and news feeds along with who is reading them and create a hub-and-spoke view of the world.

Entity extraction algorithms identify important people, places, and entities in the document and then associate them with the text in close proximity. If there is a quote from a CEO, and his or her name appears, the entity-extraction algorithm will recognize that part of the text as being ascribed to the CEO. Currently, these algorithms are used mainly for associating companies rather than individuals with text. In addition, entity extraction is more accurate than sentiment scoring, and false positives come with the territory. If an influential person on social media had a common name, for instance, then conceivably, the comments of someone with the same name could be tracked in error.

Derwent Capital, a U.K.-based hedge fund, is a pioneer in using these tools to forecast stock market movements based on social media sentiment. By looking at tweets or other entries, they gauge how people feel about the economy or particular companies.

But John Bates, chief technology officer of Progress Software, is skeptical. "By the time you analyze the sentiment, it's a trailing indicator; it's not a leading indicator. You need to analyze real-time events as well."

Complex event processing (CEP) enables firms to do that, and now some firms are interested in incorporating voice, video, and social media into their CEP strategy.

Progress Software was a pioneer in using news feeds as part of the forecasting process. Their work inspired Dow Jones to develop Elementized News Feeds, which deliver ultra-low-latency economic data and corporate news in a precisely tagged XML format. According to Dow Jones, the feeds can be parsed and embedded directly into quantitative analysis models and algorithmic trading programs. This service allows traders, quants, and strategists to make better trade decisions, cut execution times to microseconds, and perform deeper, historical trend analysis.

One approach firms can take is to implement a fusion of technology that can take in many types of media and do CEP. Around the edge of that, they can deploy specialized processing engines that can support social media, audio, and video processing to convert the unstructured data into events that can be correlated with each other.

Say a firm wants to use satellite images to track the movements of oil tankers and build this information into their oil price forecasts. The output would be an event describing an oil tanker that has been detected, and it would be exactly the same as if one were getting that information from a structured data source.

The Four Vs

When considering unstructured data, investors need to understand the four Vs: volume, velocity, variety, and veracity.

Volume is a challenge for financial firms that want to automatically feed unstructured data into their trading algorithms. Most sentences in the English language are constructed using 10,000 words. More expressively rich documents may be constructed using 100,000 words. In a 100,000-word vector, one would look to see whether specific words are present. Think of this approach as an extremely long but sparsely populated vector that represents one document inside the computer. Analysts may be working with hundreds of thousands of documents, effectively creating a huge matrix.

"We have these massive matrices with millions of rows and hundreds of thousands of columns, and they are sparsely populated," says Chid Apte, director of analytics at IBM Research. "That is the base structure on which your algorithm needs to do the analytics."

Velocity is also important because the information is only useful if analytics can be done in a timely fashion. Timely may mean different things to different people. For a financial trading application, timely may mean a sub-millisecond response. For another type of application, timely may mean as of last night. Velocity is driving technology advances in streaming analytics.

Variety and *veracity* also come into play. Unstructured data can come in the form of text, images, audio, and video. Users must grapple with the quality of data.

Only the Beginning

It is possible to snap together off-the-shelf components, listen for all tweets that include a specific company's name, and build a strategy that buys when sentiment is up and sells when sentiment is down.

"You could build that today easily, and you would almost be guaranteed to lose money," says Tibbetts. "You might as well flip coins. That's not a sophisticated-enough strategy."

In reality, firms that want to process unstructured data need a sophisticated computer infrastructure. They also require people with specialized skills in technology and optimization, as well as developing and programming algorithms.

"That doesn't scale up because we clearly have more of an appetite and need for the uses and consumption of these kinds of solutions than there are people who can build them," adds Apte. This problem is driving research into scalability and automation. Experts are trying to find ways to automate an end-to-end execution of these applications so they can be used in a repeatable fashion without necessarily requiring an expert human to be involved in delivering the answer.

There is a huge momentum around the topic of big data (i.e., leveraging unstructured data, integrating it with structured data, and then doing analytics on it in a high-volume, high-velocity world). But for now, none of the semantic strategies are as mature as volume-weighted average price (VWAP) or pairs trading.

"Next-generation data analytics will push machines closer to being able to do what humans do when they read text."

"Firms are still coming to understand what the real-time data-processing stack is," says Tibbetts. "The big focus for many people is to understand where does real time fit into my organization? What parts of my business need to be real time this year? What parts of my business need to be real time next year? Where is real time a competitive advantage?"

Once these tools mature, the next step will be to implement them across all asset classes and in the middle- and back-offices, as well as to offer cloud-based applications for mobile users. In the next decade, unstructured data analysis could become a mainstream practice. At that point, next-generation data analytics will push machines closer to being able to do what humans do when they read text.

For now, someone still has to handcraft an algorithm that gets embedded in these frameworks to do the analytics, and this method requires careful inspection, design, and experimentation. The ultimate goal is to replace handcrafting with a self-adapting, self-learning mechanism that has the intelligence to figure out dynamically whatever needs to be done to get insights from the information at hand.

IBM's Watson project, the system built to play the TV game show *Jeopardy*, is an early example of a self-adapting, self-learning system. Watson can answer questions at great speed with a high degree of accuracy, yet it is focused on a narrow domain. The question is whether through an evolutionary approach it can be broadened to serve general purposes.

"We are indeed working with a huge mass of unstructured information," says Apte. "We are dynamically self-adjusting and learning how to optimize the answering of questions that are posed to the system."

While traditional analysis will remain important, firms need to consider other approaches to forecasting. Analytics technology is becoming more sophisticated in its ability to leverage unstructured data. Those who understand what it can and cannot do and learn how to use it properly may be able to gain competitive advantage in the near term. Over the medium to long term, it could become an imperative.

Sherree DeCovny is a freelance journalist specializing in finance and technology.

RECOMMENDED RESOURCES

"China's Fragile Foundations" *CFA Institute Conference Proceedings Quarterly* (Ahead of print, 6 June 2012) (www.cfapubs.org)