Teaching Computers to Understand Human Language: How Natural Language Processing is Reshaping the World of Finance

Alex LaPlante, Research Manager; Thomas F. Coleman, Chief Research Officer

Nov. 7, 2016

From the beginning of recorded history until 2003, man created a few dozen exabytes of information. Today, this amount of information is created within days — 90% of the world's data was created in the last two years alone. [1] This surge in data generation is largely driven by the advent of web-based social networking and content-sharing services which allow users to create and share their own content, ideas, and opinions. Every minute 100,000 new tweets are shared, 571 new websites are created, and 205 million e-mails are sent.

This vast expanse of data contains valuable information that can be used to augment a wide range of processes in the financial sector including fraud detection, market prediction, and customer relationship management. Nonetheless, since the majority of this data is created for human consumption it is stored in unstructured formats, such as word documents, PDFs, social media posts, and audio files, and cannot be directly processed by computers. Structured data, like that found in relational databases, is highly organized data that can be readily searched, processed and exploited for analytic purposes using straightforward algorithms. On the other hand, unstructured data, which makes up approximately 80% of available data, requires a larger computational effort and a deep understanding of natural language by machines in order for automatic analysis to be performed. Although we have yet to reach the point where computers understand all the nuances of human language, research efforts in the field of Natural Language Processing (NLP) are bringing us ever closer to this reality.

NLP is a sub-section of the artificial intelligence domain that is focused on teaching computers to understand natural human languages. To enable understanding, computers require unambiguous and precise messages. However, human communications are often ambiguous and imprecise. NLP encompasses a range of theory-motivated computational techniques which attempt to address this disconnect by allowing for the automatic analysis and representation of human languages. It allows computers to make inferences about and provide context to language, consequently enabling them to mine valuable data from large quantities of unstructured data in a timely manner.

Unsurprisingly, a growing number of companies, including financial institutions, are now exploiting NLP to enhance their analytics frameworks and to gain a better understanding of their clients and their broader operational environments. A recent report by

MarketsandMarkets estimated that the value of the NLP market was $5.7 billion in 2015 and projected it to grow to $13.4 billion by 2020. [2] As NLP continues to increase in popularity over the coming years, institutions that do not employ NLP techniques to exploit the masses of available unstructured data will be at a large competitive disadvantage and will miss out on actionable insights and their associated, potentially significant, monetary gains.

To shed light on the emerging role of NLP in finance, this report will provide a brief introduction to NLP and will detail how financial institutions can employ these techniques to extract valuable information and improve their risk management processes.

Syntactics, Semantics and Pragmatics

Language is inherently complex and relies not only on denotative knowledge but also on our ability to understand connotation and context. Not surprisingly, developing computational methods to capture all of the nuances of human language becomes increasingly difficult as we move from syntax-based approaches to those that consider semantics and pragmatics. Syntactic NLP, which takes advantage of arbitrary key words, punctuation, and word frequencies, is the most commonly employed approach for information retrieval and extraction, auto-categorization and topic modelling.[3] Although these word-based techniques have worked relatively well thus far, they are not able to extract and manipulate textual meaning and are thus susceptible to deceptive data. For example, consider the word “accident”. In many contexts this word is viewed in a negative light: “I got in an accident”. Of course, that is not always the case: “I met my wife by accident”. Syntactic methods that consider only the valence (positive or negative) and frequency of keywords may identify both examples as being negative. While more advanced statistical syntactic NLP methodologies take into account punctuation, word occurrence frequencies, and the valence of keywords as well as other arbitrary words, they are still generally semantically weak.

Improving upon the shortcomings of syntactic methods, semantic NLP focuses on capturing the intrinsic meaning associated with natural language. These methods move away from the blind-usage of key words and word frequencies and rely instead on the denotative and connotative information contained within language. To achieve this, semantic NLP can exploit external knowledge, such as taxonomies or semantic knowledge bases, or can rely purely on the inherent semantics of the set of documents being analyzed. Although semantic NLP brings us one step closer to achieving natural language understanding, it does not recognize narratives or context, two key aspects of human reasoning and decision making. Pragmatic NLP, which is of great interest to AI researchers and seen as the future of NLP, looks to assess semantics in time, performing parallel and dynamic comparisons given different contexts and with respect to different actors. [4] While great strides have been made in semantic and pragmatic NLP, there is still some way to go before natural language understanding is truly achieved.

Computation and Modeling

Prior to the 1980s, the majority of NLP analyses were performed manually based on complex sets of hand-written rules. As computational power and capabilities increased in the mid-80s, studies using computers and simple machine learning algorithms to perform basic text analysis began to emerge. [5] Many of the early computational NLP techniques were based on a system of hard if-then rules which were comparable to the existing hand-written rules. Overtime research has progressively moved from these traditional non-probabilistic methods towards statistical models which make probabilistic decisions based on the inputted data.

NLP models can be developed to analyze data obtained from a wide range of web and print sources, including newspapers, reports, e-mails, blogs, and twitter feeds. Depending on the type of data available for model development, supervised, semi-supervised or unsupervised learning algorithms can be used. Supervised learning requires labelled training data that is comprised of input vectors and their corresponding desired output. On the other hand, unsupervised and semi-supervised algorithms are able to learn from unlabelled or partially labelled input data. Unsupervised and semi-supervised learning is more difficult and typically less accurate than supervised learning for a given amount of data. That being said, these methods offer the advantage of exploiting the massive amounts of non-labelled data available today.

NLP in Finance

Financial institutions have access to many proprietary and non-proprietary data sources that can be used not only for day-to-day operations and reporting but can also be mined for actionable insights that can lead to improved business practices and monetary gains. Increasingly, NLP techniques are being employed to extract value from the largely untapped sea of unstructured data available to financial institutions, allowing for enhancements to a wide range of processes across many lines of business.

NLP has proved to be a useful tool for information retrieval and the classification of financial statement content. Studies have shown that NLP can be used to significantly reduce the manual processing required to retrieve corporate data from sources including the Security Exchange Commission's (SEC's) EDGAR (Electronic Data Gathering, Analysis, and Retrieval) database, financial reports, press releases, and news articles. [6, 7] Moreover, NLP can be used to verify the consistency between company reports and financial statements. [8, 9] Given this ability to quickly access and verify relevant, filtered information, financial analysts are able to provide more comprehensive and informative reports upon which management can base their decisions. An example of this is IBM's AlchemyData News API which can create real-time alerts based on various news articles including analysts' reports. NLP can also be used to improve internal reporting efforts and can provide timely updates on key matters.

NLP provides an efficient means of monitoring consumer and investor sentiments. By applying NLP-based sentiment analysis techniques to news articles, reports, and social media or other web content, one can effectively determine whether those sources have a positive or negative tone. This can be used to monitor client sentiments about a firm, investor sentiments about the market, or the sentiment contained within financial reports.

Using NLP in conjunction with other analytical methods can also aid in the prediction and detection of fraud. Deceptive statements often contain certain language patterns, like increased usage of negative-emotion words and reduced usage of first-person pronouns. [10] NLP can identify and exploit these patterns to detect fraudulent statements in corporate communications like annual securities filings, e-mails, and even transcripts of analyst conference calls. [11, 12, 13, 14] For example, Lloyd's Banking Group has employed NLP in conjunction with machine learning techniques to identify fraudulent phone calls. Similarly, these techniques can be used by financial institutions and regulators to scan message boards and e-mails, and track regulatory filings and trade data to help identify possible illegal insider trading.

Perhaps the most noted application of NLP is its use in the prediction of stock prices and other securities market activity. Deutsche bank has shown that the use of NLP-based techniques provides significant improvement to quantitative investing models and stock price prediction. [15, 16] Similarly, financial institutions can exploit these predictive capabilities to estimate profitability, default risk, and other measures of an external firm's performance. More broadly, financial institutions can apply NLP techniques to their large banks of unstructured data to identify new predictive variables that can be used to augment analytics processes across all lines of business.

NLP and Risk

Given its extensive applicability, NLP has large implications for a financial institution's risk management practices. As previously noted, NLP can be used to reduce model risk through improved risk models (ex. default models) and can provide enhanced fraud and insider trading detection. Beyond that, NLP can greatly aid in ensuring regulatory and legal compliance and can facilitate communications with regulators by aggregating relevant data from different lines of business. NLP’s effective extraction of metadata and “understanding” of content allows one to efficiently track changes to regulatory requirements and determine compliance related costs. As a result, NLP can provide the fundamental information required for financial institutions to take a risk-based view of regulatory compliance. Similarly, this efficient data aggregation can greatly reduce the resources required for key risk reporting and auditing processes.

As the adoption rate of NLP-based technologies continues to grow, the broader operational environment of financial institutions and the expectations of their client bases will shift. It will become ever more difficult to remain competitive with firms who take advantage of the cost savings and improved operational efficiency that NLP can provide. In managing their operational risks, financial institutions will need to assess the opportunities that NLP presents and weigh the risks of early, late or non-adoption.

What’s next?

Although there is still a ways to go before full language understanding is achieved, the field of NLP has made significant progress, allowing for technologies that have and will continue to revolutionize how financial institutions operate. In the near term, we are likely to see an increase in NLP-based technologies like chatbots and robo-advisers. In the New Year, Bank of America will be launching “Erica,” a virtual assistant who you can chat with via audio or text and who will analyze client accounts and answer questions. As advancements allow us to move away from syntactic NLP towards semantic and even pragmatic NLP, language understanding and sentence generation will only become more accurate and will allow computers to closely mimic real human interaction. Scientists and researchers are exploring a plethora of NLP-based technologies that will reshape how we interact with other people and the broader world around us. Microsoft’s Project Tokyo, for example, aims to deliver AI-based technologies that will help the visually impaired navigate their social, physical, and textual environments. Other NLP-based advancements may come in the form of personal real-time universal translators, which will drastically impact international business practices, or “zero UI” (user interface) platforms, which will eliminate the need to use mice or touchscreens to interact with our computers and mobile devices. At this point there is really no telling where these technologies will lead us!

Technological advancements over the last few decades have prompted the exponential creation of data in textual, audio, and visual formats. These vast data sources contain potentially useful information that, if properly extracted, can be used to augment business practices and provide valuable insights. Natural language processing allows computers to mine pertinent information from large quantities of unstructured data far more efficiently than would be possible using manual techniques. Unsurprisingly, leading innovators and tech giants like Google and Facebook rely heavily on NLP techniques along with other AI methodologies to glean cutting-edge insights from their data and provide dynamic customer experiences. Given the data-driven and customer-centric nature of financial institutions, NLP presents countless opportunities to create value through enhanced business practices. As the scope of unstructured data collection continues to grow, and the adoption and understanding of NLP methodologies expands, financial institutions that fail to exploit this technology will increasingly find themselves at a competitive disadvantage. Institutions that invest in NLP, on the other hand, will find unprecedented insights and opportunities.

References

[1] IBM, (2016) "Bringing Big Data to Enterprise", cited: July 30, 2016, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[2] Markets and Markets, (2016) "Natural Language Processing by Market Type, Technologies, Development, Vertical and Region"

[3] Cambria, E., White, B., (2014) "Jumping NLP Curves: A Review of Natural Language Processing Research", *IEEE Computational Intelligence Magazine,* May 2014

[4] Cambria, E., Howard, N., (2013) "Intention Awareness: Improving Upon Situation Awareness in Human-Centric Environments", *Human-Centric Computing Information Sciences,* vol.3, no.9

[5] Fisher, I., Garnsey, M., Hughes, M., (2016) "Natural Language Processing in Accounting, Auditing and Finance: A Synthesis of the Literature with a Roadmap for Future Research", *Intelligent Systems in Accounting, Finance and Management*, vol. 23, issue 3, pp. 157-214

[6] Gerde, J., (2003) "EDGAR-Analyzer: Automating the Analysis of Corporate Data Contained by the SEC's EDGAR Database", *Decision Support Systems*, vol. 35, no. 1, pp. 7-29

[7] Grant, G. H., Conlon, S. J., (2006) "EDGAR Extraction System: An Approach to Analyze Employee Stock Option Disclosures", *Journal of Information Systems*, vol. 20, no. 2, pp. 119-142

[8] Back, B., Toivonen, J., Vanharanta H., Visa, A., (2001) "Comparing Numerical Data and Text Information from Annual Reports using Self-Organizing Maps", *International Journal of Accounting Information Systems*, vol. 2, no. 4, pp. 249-269

[9] Chen, C. L., Lui, C. L., Change, Y. C., Tsai, H. P., (2013) "Opinion Mining for Relating Subjective Expressions and Annual Earnings in US Financial Statements", *Journal of Information Science and Engineering, vol. 29, no. 4, pp. 743-764*

[10] Keila, P. S., Skillicorn, D. B., (2005) "Detecting Unusual and Deceptive Communication in E-mail", *CASCON '05 Proceedings of the 2005 Conference of the Center for Advanced Studies on Collaborative Research,* IBM Press, pp. 17-20

[11] Goel, S., Gangoly, J., Faerman, S., Uzuner O., (2010) "Can Linguistic Predictors Detect Fraudulent Financial Filings?" *Journal of Emerging Technologies in Accounting*, vol. 7, no. 1, pp. 25-46

[12] Larcker, D. F., Zakolyukina, A. A., (2012) " Detecting Deceptive Discussions in Conference Calls", *Journal of Accounting Research*, vol. 50, no. 2, pp. 495-540

[13] Cecchini, M., Aytug, H., Koehler, G. J., Pathak, P., (2010) "Making Words Work: Using Financial Text as a Predictor of Financial Events", *Decision Support Systems*, vol. 50, no. 1, pp. 164-175

[14] Purda, L., Skillicorn, D., (2014) " Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection", *Contemporary Accounting Research*, vol. 32, no. 3, pp. 1193-1223

[15] Cahan, R., Luo, Y., Jussa, J., Alvarez, M., (2010) "Signal Processing: Beyond the Headlines", *Deutsche Bank*

[16] Cahan, R., Luo, Y., Jussa, J., Alvarez, M., Chen, Z., Wang, S., (2011) "Signal Processing: Quant 2.0", *Deutsche Bank*