TEXTUAL-ANALYSIS FOR RESEARCH IN PROFESSIONAL JUDGMENT AND DECISION MAKING, AUDIT AND ASSURANCE, RISK, CONTROL, GOVERNANCE, AND REGULATION

Editorial

Data are more than numbers. It also includes textual communications, reports, social media posts, images, spoken word data and videos. Accounting and audit research increasingly analyzes textual data in many different ways to investigate a wide range of research questions (Fisher, 2018; Fisher *et al.*, 2016; Loughran and McDonald, 2016; Teoh, 2018; Vasarhelyi *et al.*, 2015; Zhang *et al.*, 2019), as illustrated by the following four studies in this eBook:

- Chang and Stone (2019a) examine the association between auditor selection and audit proposal readability, measuring readability using a simple linear combination of counts of syllables, words and sentences in the proposals;
- Chang and Stone (2019b) study the association between audit firm size and report readability using natural language processing (NLP) measures that are new to audit literature;
- Boskou *et al.* (2019) use supervised machine learning to develop a proxy for internal audit quality based on textual disclosures extracted from management's discussion and analysis (MD&A); and
- Zorio-Grima and Carmona (2019) use both manual content analysis and a variety of automated textual analysis approach to study transparency reports issued by Big 4 audit firms.

These studies demonstrate that quantitative approaches may be used to derive insights from textual data once the text has been appropriately transformed. Machine learning and data visualization tools commonly associated with "Big data" studies are frequently used for textual analysis, even when sample sizes are small. As discussed below, the four articles in this eBook lay bare both the opportunities for adopting text analytic techniques and the challenges of preparing textual data and understanding the models used.

Data transformation

Once textual data has been parsed, filtered and structured, it may be analyzed with techniques and software tools commonly used to analyze numeric data. For example, the machine learning algorithms used by Boskou *et al.* (2019) to classify internal control quality based on textual disclosures include algorithms used in financial restatement, going concern and financial statement fraud studies that use numerical data (Chye Koh and Kee Low, 2004; Dutta *et al.*, 2017; Kirkos *et al.*, 2007; Perols, 2011). All four articles in this issue detail the transformation

This paper forms part of a special section "Textual-Analysis for Research in Professional Judgment and Decision Making, Audit and Assurance, Risk, Control, Governance, and Regulation", guest edited by Louise Hayes.

Managerial Auditing Journal
Vol. 34 No. 8, 2019
pp. 865-870

0268-6902
DOI 10.1108/MAJ-09-2019-019

(preprocessing) actions commonly used so that textual data may be analyzed quantitatively. These actions include extracting relevant portions of text (e.g. ignoring diagrams, charts and boilerplate language) and then applying additional preprocessing steps that depend upon the analytic methods and tools used. For example, in this issue, Boskou *et al.* (2019) and Zorio-Grima and Carmona (2019) describe transforming text into a matrix that can be analyzed using statistical software and machine learning techniques. Transformations used in other studies range from straight-forward keyword or dictionary-based word counts in disclosure studies (Campbell *et al.*, 2014) to vector-based measures in studies that compare document similarity (Brown and Tucker, 2011).

The success of analytics depends on data quality. Both the integrity of the data and the extent of data cleaning required to deal with outliers and missing data depend on the data source. Sources of textual data such as auditors' reports, MD&A, and news articles from reputable journals will require less cleaning effort than customer social media comments (Applebaum *et al.*, 2017). Loughran and McDonald (2016, p. 1192) elaborate on the challenges of transforming textual data for quantitative analysis and caution that "we must be careful that the imprecision of the method does not overwhelm any hoped-for gains in identifying meaning." Zorio-Grima and Carmona (2019) demonstrate that both automated textual analytics and manual content analysis may be used in the same paper to address this challenge, i.e. they adopt a "mixed, partially automated content analysis" approach that is used in many textual analysis papers (Zhang *et al.*, 2019, p. 151).

Democratization of textual analytics

Many of the analytic methods used today were developed over half a century ago (CPA/AICPA, 2018; Loughran and McDonald, 2016). Today, the increase in computing power and reduced cost of storage have led to the development of software to clean, structure, and analyze data on desktop, laptop, distributed computing and cloud computing platforms. For example, NLP software tools automate many of the preprocessing steps, although word lists used by some NLP software to parse and filter text (e.g. lists of common words such as "and" and "the" that are automatically removed and lists used for part-of-speech tagging) may need to be tailored (Guan *et al.*, 2018). Consistent with these shifts in technology, Zhang *et al.* (2019) report that there has been a shift away from traditional content analysis to more sophisticated machine learning text analytic approaches. It is noteworthy that some machine learning algorithms (e.g. multi-layered neural networks) used to model non-linear relationships (Chye Koh and Kee Low, 2004) may require little data preprocessing (Issa *et al.*, 2016) as these complex, non-statistical methods are less sensitive to outliers, correlated variables, and data sets with fewer observations than variables, i.e. data distributions commonly encountered when working with textual data.

Studies presented in this eBook use software tools to analyze textual data. Boskou *et al.* (2019) chose RapidMiner Studio, Chang and Stone (2019b) Coh-Metrix (a computational linguistic analysis tool) and Zorio-Grima and Carmona (2019) text mining and latent semantic analysis packages developed for R. While RapidMiner Studio and R run on personal computers, the Coh-Metrix software is a web-based tool accessed via the internet. Many NLP and machine learning data analytic studies use tools such as R and Python that require programming skills. However, consistent with the evolution of other in-demand technologies, fewer programming skills are required as analytic tools improve (Alles, 2015). The analytic packages available for Python and R are rapidly evolving. So are commercial vendors' connections to these open source tools and other cloud-based text analytics tools. Today, more and more solutions are being built with less help from data scientists (www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo). Given the fast-paced changes in the analytic tools

landscape, time is well spent reviewing current software options, as software tools' functionality and ease of use vary widely.

Big data techniques and small sample studies

"Big data techniques can also be applied to traditional, smaller data sets to gain additional insights" (Gepp *et al.*, 2018, p. 108). This is particularly relevant for textual analysis, as a small collection of documents is also a big collection of words and phrases (Zhang *et al.*, 2019). Thus, data analytics techniques used to search for patterns, make predictions, and draw insights from larger data sets may prove useful for analyzing small samples of textual data. In this issue, machine learning algorithms were used to analyze 28 transparency reports (Zorio-Grima and Carmona, 2019) and data from 133 publicly listed Greek companies (Boskou *et al.*, 2019). Issa *et al.* (2016) provide examples of ways textual analysis is being used today (i.e. for analyzing contracts and reviewing documents), cites recent textual analysis audit research, and discusses how textual data might be combined with data extracted from accounting information systems to automate audit tasks in the future.

Understanding and explaining models

Researchers need to understand the intuition underlying the models they adopt, including machine learning models that show increasing promise for audit analytics (Applebaum *et al.*, 2017). Depending upon the accounting and audit literature survey, the greatest proportion of automated textual analysis studies are readability studies (Fisher, 2018) or sentiment studies (Zhang *et al.*, 2019). Most of the readability studies use simple, easy-to-understand formulae based on counts of the number of words in sentences and syllables in words (Chang and Stone, 2019a; Lehavy *et al.*, 2011; Li, 2008). Similarly, sentiment studies rely on counts of occurrences of words on lists associated with certain emotions. In contrast, other analytic models are complex. In this issue, Chang and Stone (2019b) use computational linguistic techniques to measure the co-relations of words, sentences, and paragraphs on multi-dimensions and Boskou *et al.* (2019) use algorithms developed using machine learning. Frequently text analytic studies use multiple models because, as Boskou *et al.* (2019) explain, the incremental costs of running multiple models once data has been cleaned and structured are small. When multiple models are used, model choice should include considerations of not only model fit but also model interpretability (Hall *et al.*, 2017). Those who do not understand the intuition underlying models may draw unsupported conclusions from complex analytics reported by easy-to-use analytic software. The articles in this eBook explain the intuition underlying some complex machine learning algorithms (Boskou *et al.*, 2019; Zorio-Grima and Carmona, 2019) and NLP techniques (Chang and Stone (2019a, 2019b)).

Biases

Algorithms developed by artificial intelligence (AI) will be biased if based on non-representative data. Furthermore, decision-makers bring cognitive biases (e.g. anchoring, availability and confirmation bias) to the interpretation of model results (Huerta and Jensen, 2017). Microsoft recognizes the risk of bias in AI processes and warned of this in its annual SEC filing in August 2018:

> AI algorithms may be flawed. Data sets may be insufficient or contain biased information. Inappropriate or controversial data practices by Microsoft or others could impair the acceptance of AI solutions. These deficiencies could undermine the decisions, predictions or analysis AI applications produce, subjecting us to competitive harm, legal liability and brand or reputational harm (Simonite, 2019).

When adopting text analytic approaches that rely on AI processes, changes in business processes and societal trends that affect data distributions should be considered. Disclosures and language usage change over time.

Data visualization

A picture is worth a thousand words. Visualizations, along with descriptive statistics, are used to explore data, reveal patterns and identify data items needing attention. For most textual analysis projects, scanning data manually for outliers, errors and missing data are impractical. Visualization software can reduce multi-dimensional data to a visual image that humans can intuitively grasp. For example, word clouds can depict word frequencies (Zorio-Grima and Carmona, 2019) and the correlation of hundreds of words can be displayed in two dimensions on a network graph by showing their correlations as the distance between data points that represent different words, the size of which shows the word frequencies (Hall *et al.*, 2017).

In contrast, to the straight-forward visualizations produced by spreadsheet software (i.e. bar graphs, trend lines, etc.), machine learning models are frequently needed to reduce a high-dimension text data sets to two- or three-dimension visualizations. In their seminal article, Hinton and Salakhutdinov (2006, p. 507) declare that models that reduce dimensionality to facilitate visualization have:

> Been obvious since the 1980s [. . .] provided that computers were fast enough, data sets were big enough and the initial weights were close enough to a good solution. All three conditions are now satisfied.

Accordingly, visualizations such as those used by Zorio-Grima and Carmona (2019) to illustrate the relationships between narrative disclosures in transparency reports are beginning to appear in audit and accounting literature. Data visualizations need to be carefully prepared and used cautiously as decision-makers draw conclusions swiftly from images and are subject to cognitive biases that affect their interpretation of visualizations (Dilla *et al.*, 2010; Rose *et al.*, 2016).

Conclusion

Adoption of automated textual analysis is accelerating as data analytic skills improve and easier-to-use data preparation and analysis tools become more readily available. The articles in this eBook not only contribute to audit literature but also demonstrate the promise of text analytics for future research. However, there is a risk that the inferences and conclusions drawn from text analytics may be unfounded when these tools are used by those unfamiliar with the challenges of preprocessing textual data, evaluating and explaining advanced models, and interpreting data visualizations. Collectively, the articles in this eBook invite the reader to make greater and more informed use of textual analysis in the future.

Louise Hayes
Department of Management, University of Guelph, Guelph, Canada

References

Alles, M.G. (2015), "Drivers of the use and facilitators and obstacles of the evolution of big data by the audit profession", *Accounting Horizons*, Vol. 29 No. 2, pp. 439-449.

Applebaum, D., Kogan, A. and Vasarhelyi, M.A. (2017), "Big data and analytics in the modern audit engagement: research needs", *Auditing: A Journal of Practice and Theory*, Vol. 36 No. 4, pp. 1-27.

Boskou, G., Kirkos, E. and Spathis, C. (2019), "Classifying internal audit quality using textual analysis: the case of auditor selection", *Managerial Auditing Journal*, available at: https://doi.org/10.1108/MAJ-01-2018-1785

Brown, S.V. and Tucker, J.W. (2011), "Large-Sample evidence on firms' year-over-year MD&a modifications", *Journal of Accounting Research*, Vol. 49 No. 2, pp. 309-346.

Campbell, J.L., Chen, H., Dhaliwal, D.S., Lu, H. and Steele, L.B. (2014), "The information content of mandatory risk factor disclosures in corporate filings", *Review of Accounting Studies*, Vol. 19 No. 1, pp. 396-455.

Chang, Y.-T. and Stone, D. (2019a), "Proposal readability, audit firm size and engagement success: do more readable proposals win governmental audit engagements?", *Managerial Auditing Journal*, available at: https://doi.org/10.1108/MAJ-01-2018-1665

Chang, Y.-T. and Stone, D. (2019b), "Why does decomposed audit proposal readability differ by audit firm size? A Coh-Metrix approach", *Managerial Auditing Journal*, available at: https://doi.org/10.1108/MAJ-02-2018-1789

Chartered Professional Accountants of Canada (CPA Canada) and the American Institute of CPAs (AICPA) (2018), "A CPA's introduction to AI: from algorithms to deep learning, what you need to know", available at: www.cpacanada.ca/en/business-and-accounting-resources/other-general-business-topics/information-management-and-technology/publications/a-cpa-introduction-to-ai (accessed 28 July 2019).

Chye Koh, H. and Kee Low, C. (2004), "Going concern prediction using data mining techniques", *Managerial Auditing Journal*, Vol. 19 No. 3, pp. 462-476.

Dilla, W., Janvrin, D.J. and Raschke, R. (2010), "Interactive data visualization: new directions for accounting information systems research", *Journal of Information Systems*, Vol. 24 No. 2, pp. 1-37.

Dutta, I., Dutta, S. and Raahemi, B. (2017), "Detecting financial restatements using data mining techniques", *Expert Systems with Applications*, Vol. 90, pp. 374-393.

Fisher, I.E. (2018), "A perspective on textual analysis in accounting", *Journal of Emerging Technologies in Accounting*, Vol. 15 No. 2, pp. 11-13.

Fisher, I.E., Garnsey, M.R. and Hughes, M.E. (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 No. 3, pp. 157-214.

Gepp, A., Linnenluecke, M.K., O'Neill, T.J. and Smith, T. (2018), "Big data techniques in auditing research and practice: current trends and future opportunities", *Journal of Accounting Literature*, Vol. 40, pp. 102-115.

Guan, J., Levitan, A.S. and Goyal, S. (2018), "Text mining using latent semantic analysis: an illustration through examination of 30 years of research at JIS", *Journal of Information Systems*, Vol. 32 No. 1, pp. 67-86.

Hall, P. Phan, W. and Ambati, S. (2017), "Ideas on interpreting machine learning: mix-and-match approaches for visualizing data and interpreting machine learning models and results", O'Reilly Media blog March 15, 2017, available at: www.oreilly.com/ideas/ideas-on-interpreting-machine-learning (accessed 28 July 2019).

Hinton, G.E. and Salakhutdinov, R.R. (2006), "Reducing the dimensionality of data with neural networks", *Science (New York,* NY*)*, Vol. 313 No. 5786, pp. 504-507.

Issa, H., Sun, T. and Vasarhelyi, M.A. (2016), "Research ideas for artificial intelligence in auditing: the formalization of audit and workforce supplementation", *Journal of Emerging Technologies in Accounting*, Vol. 13 No. 2, pp. 1-20.

Kirkos, E., Spathis, C. and Manolopoulos, Y. (2007), "Data mining techniques for the detection of fraudulent financial statements", *Expert Systems with Applications*, Vol. 32 No. 4, pp. 995-1003.

Lehavy, R., Li, F. and Merkley, K. (2011), "The effect of annual report readability on analyst following and the properties of their earnings forecasts", *The Accounting Review*, Vol. 86 No. 3, pp. 1087-1115.

Li, F. (2008), "Annual report readability, current earnings, and earnings persistence", *Journal of Accounting and Economics*, Vol. 45 Nos 2/3, pp. 221-247.

Loughran, T. and McDonald, B. (2016), "Textual analysis in accounting and finance: a survey", *Journal of Accounting Research*, Vol. 54 No. 4, pp. 1187-1230.

Perols, J. (2011), "Financial statement fraud detection: an analysis of statistical and machine learning algorithms", *Auditing: A Journal of Practice and Theory*, Vol. 30 No. 2, pp. 19-50.

Simonite, T. (2019), "Google and Microsoft warn that AI may do dumb things", Wired, available at: www.wired.com/story/google-microsoft-warn-ai-may-do-dumb-things/ (accessed August 11 2019).

Teoh, S.H. (2018), "The promise and challenges of new datasets for accounting research", *Accounting, Organizations and Society*, Vols 68/69, pp. 109-117.

Vasarhelyi, M.A., Kogan, A. and Tuttle, B.M. (2015), "Big data in accounting: an overview", *Accounting Horizons*, Vol. 29 No. 2, pp. 397-407.

Zhang, M.C., Stone, D.N. and Xie, H. (2019), "Text data sources in archival accounting research: Insights and strategies for accounting systems' scholars", *Journal of Information Systems*, Vol. 33 No. 1, pp. 145-180.

Zorio-Grima, A. and Carmona, P. (2019), "Narratives of the big-4 transparency reports: country effects or firm strategy?", *Managerial Auditing Journal*, available at: https://doi.org/10.1108/MAJ-09-2018-1994

Further reading

Rose, A., Rose, J., Sanderson, K. and Thibodeau, J. (2017), "When should audit firms introduce analyses of big data into the audit process?", *Journal of Information Systems*, Vol. 31 No. 3, pp. 81-99.