When big data made the headlines: mining the text of big data coverage in the news media

Murtaza Haider

Ted Rogers School of Management,
Ryerson University,
350 Victoria Street, Toronto
ON M5B 2K3, Canada
Email: murtaza.haider@ryerson.ca

Amir Gandomi*

Frank G. Zarb School of Business,
Hofstra University,
1000 Hempstead Turnpike,
Hempstead, NY 11549, USA
Email: amir.gandomi@hofstra.edu
*Corresponding author

Abstract: Big data-driven analytics emerged as one of the most sought-after business strategies of the decade. This paper reviews the news coverage of this phenomenon in the popular press. The study uses natural language processing (NLP) and text mining algorithms to determine the focus and tenor of the news media reporting of big data. A detailed content analysis of a five million-word corpus reveals that most news coverage focused on the newness of big data technologies that showcased usual suspects in big data geographies and industries. The insights gained from the text analysis show that big data news coverage indeed evolved where the initial focus on the promise of big data moderated over time. This study also offers a detailed exposé of text mining and NLP algorithms and illustrates their application in news content analysis.

Keywords: big data; news content analysis; text mining; natural language processing; NLP; topic modelling; modal verb analysis.

Reference to this paper should be made as follows: Haider, M. and Gandomi, A. (2021) 'When big data made the headlines: mining the text of big data coverage in the news media', *Int. J. Services Technology and Management*, Vol. 27, Nos. 1/2, pp.23–50.

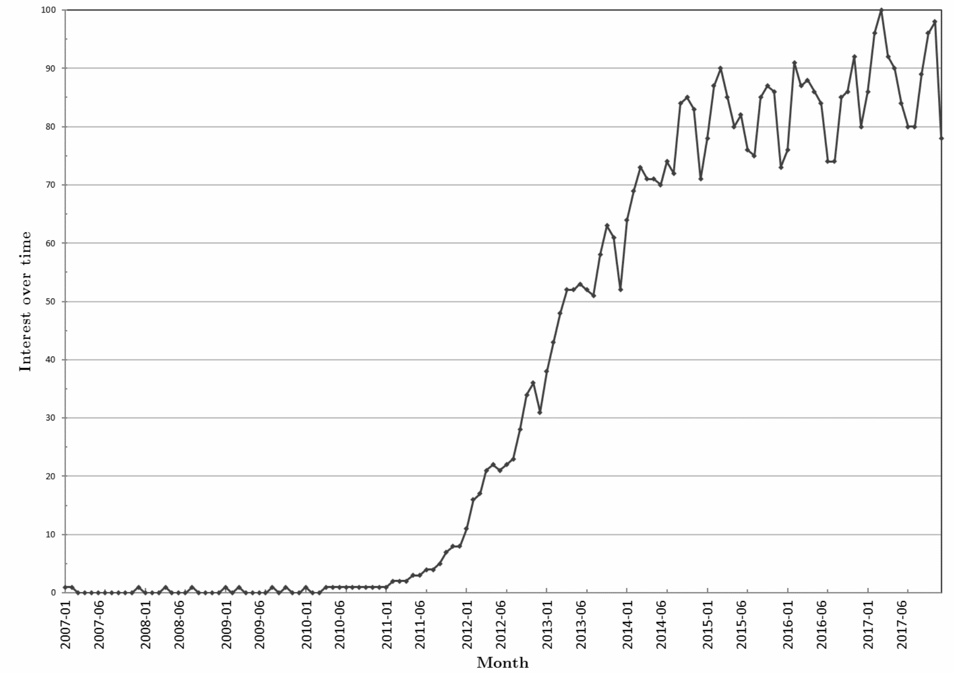
Biographical notes: Murtaza Haider is a Professor of Real Estate Management at Ryerson University. He also serves as the Research Director of the Urban Analytics Institute. He holds an Adjunct Professorship of Engineering at McGill University in Montreal. His research interests include business analytics, data science, housing market dynamics, transport/infrastructure/urban planning, and human development in Canada and South Asia.

Amir Gandomi is an Assistant Professor in the Department of Information Systems and Business Analytics at Frank G. Zarb School of Business. His research interests include optimisation (convex and non-convex), machine learning, reinforcement learning, and natural language processing with a focus on applications in healthcare and marketing.

1 Introduction

In the fast-evolving world of technological innovation, big data analytics are unique. Big data, as a business strategy (and not necessarily as an innovative technology), appeared on the horizon only in 2011 and rapidly became one of the leading technological and business trends (Gandomi and Haider, 2015). From just 229 news stories about 'big data' published in major news and business publications before 2011 to over 35,000 stories from 2011 to 2015, the growth in big data popularity is phenomenal.

Figure 1 Google trends searches for 'big data' from 2007 to 2017 (an interest value of 100 represents the peak popularity during this period)



Big data's growing fame has been complemented by its rapid deployment in almost all industrial sectors (Begdache et al., 2018a, 2018b; Janssen et al., 2017). Business executives and managers were quick to respond to the emerging trends as they procured hardware and software solutions followed by a global hunt for data scientists (Agarwal and Brem, 2015). Industry observers were quick to highlight the impending shortage in human capital needed to enable big data and analytics-driven solutions for large and small enterprises (Manyika et al., 2011). While big data continues to be in vogue, some veteran analysts are of the view that the concept might have reached a saturation point (Vanian, 2016). They cite a slowing growth trend in Google searches for big data as evidence (see Figure 1).

Given the short timespan in which big data rose to prominence, the coverage of big data technologies and applications has lagged in the academic press (Gandomi and Haider, 2015). While the popular press published over 86,700 news items during 2011 and 2017 (Dow Jones Factiva), the academic press generated much fewer publications during the same time. Also, most academic papers appeared in journals focused on computer science (84%) and engineering (23%), which are not necessarily directed at business readers. Only 5% of the academic papers on big data appeared in academic publications focusing on business and economics. Because a paper could be counted in multiple subject areas, the percentages, when summed, are greater than 100% (Table 1).

Table 1 Academic literature on big data in various disciplines (articles with 'big data' in their title published during 2011 to 2017 searched using the Web of Science research index)

Discipline	*Counts*	*Percentage*
Computer science	8,641	84.4%
Engineering electrical electronic	2,375	23.2%
Telecommunications	955	9.3%
Management	317	3.1%
Information science library science	303	3.0%
Operations research management science	255	2.5%
Business	239	2.3%
Economics	166	1.6%

The big data spend has been equally significant. Industry reports reveal that as soon as big data appeared on the horizon, investment in software and hardware followed. Gartner reported that business analytics commanded $12-billion in spending in 2011 (Kristal, 2012). Wikibon, an open source knowledge sharing forum, reported that big data spend was $19.6 billion in 2013, rising to $27.4 billion in 2014 (Kelly, 2015). International Data Corporation (IDC) estimates the market for big data technology and services will reach $48.6 billion by 2019 (IDC, 2015). Ultimately, big data spending is estimated to reach $84 billion by 2026 (Columbus, 2015).

The sustained increase in big data investments raises one important question: What sources of information (knowledge) business executives relied upon for their procurement decisions about big data in the early days? Could it be true that the popular press (news media), which was quick to realise the emerging tech trends and reported on it in real time, was the primary source of information for those responsible for implementing big data solutions?

If the news media and other sources such as technical blogs were the dominant sources of information, it becomes necessary to explore the contents of big data coverage in the news media for their objectivity. The earlier, and even subsequent, coverage of big data could have contributed to the hype where the coverage focused primarily on the oft-cited stories of big data technologies generated by the likes of McKinsey Global Institute while ignoring the inherent limitations of big data, such as correlation does not imply causation and that with hundreds of millions of observations, one is likely to find statistically significant correlation among variables.

This paper accomplishes the following tasks. Relying on the published literature, the paper shows that business executives and leaders are more likely to consult news media and other non-academic literature for information about new and emerging trends. In the absence of timely big data coverage in the academic press that focussed on business and economics, and the extensive coverage of the same in the popular press, it is likely that the popular press could have had a far-reaching impact on opinions formed about the efficacy of big data. Therefore, this paper undertakes a systematic content analysis of the news media during 2010 and 2017.

The content analysis determines whether the news coverage essentially served as advertorial by highlighting in the editorial what the major hardware and software vendors proclaimed in their advertisements. Such coverage would emphasise the promise of big data, rather than report on the realised benefits and costs of implementing big data technology. By adopting natural language processing (NLP) techniques, this paper introduces a method to investigate the tenor and tone of the big data narrative and the possible promotional content. This paper also deploys text mining algorithms to determine the primary themes covered in the news.

The rest of the paper is organised as follows. Section 2 reviews the related literature. Sections 3 details the methods used to collect, pre-process, and analyse the data. Section 4 presents the results. Section 5 provides a summary of the findings and contributions.

2 Literature review

The literature review comprises two sections. The first section presents a formal definition of big data and explains its origins as a marketing concept. The second section focuses on the information sources business executives rely upon to learn about the new technological trends.

2.1 Big data

A caveat is in order at the onset. This paper does not contend that the academic literature on the storage and analysis of big data, scalable algorithms, and efficient computing has been missing. The academic press has been the leading source of independent and critical evaluation of such topics. This paper instead focuses on big data as a marketing concept that helped popularise big data beyond computing, IT, and analytics domains.

Big data as an expression emerged as early as in March 1996 in a Silicon Graphics (SGI) advertisement in the *Black Enterprise* magazine (Diebold, 2012). Subsequent SGI advertisements in *Info World* and *CIO* magazines referred to big data and a 1998 slide deck produced by SGI's John Mashey entitled 'Big Data and the Next Wave of InfraStress' also helped popularise the term (Diebold, 2012; Fan and Bifet, 2013). Big data initially gained traction in academic circles in a book titled *Predictive Data Mining: A Practical Guide* where the authors warned: "in theory, big data can lead to much stronger conclusions for data-mining applications, but in practice, many difficulties arise" (Weiss and Indurkhya, 1997).

The first academic paper to mention big data was Francis X. Diebold's 2000 article 'Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting' (Fan and Bifet, 2013). In 2001, Laney made a significant contribution to the growing big data literature through an unpublished research note for the Meta Group entitled '3-D Data Management: Controlling Data Volume, Velocity, and Variety'. While making no explicit reference to the term 'big data', this article nonetheless introduced the concept of the 'three Vs' that is still used widely in big data research. *Volume* refers to the total amount of data generated from all sources, *velocity* refers to the speed of data transfer, and *variety* refers to the different types of data (e.g., audio, video, image, text) being collected (Gandomi and Haider, 2015). Subsequent research has extended the 3V concept to include additional Vs such as *value* (discovering hidden advantages that aid in the decision-making process), *variability* (changes in data structure and interpretation), and *veracity* (examining trust and uncertainty in data capture, storage, and analysis) (Fan and Bifet, 2013; Ward and Barker, 2013; Sivarajah et al., 2017).

In the past few years, big data has become a field of research and discipline unto itself just as Diebold (2012) foreboded: "big data is at the heart of modern science and business". It is being applied in a wide variety of disciplines including, but not limited to healthcare, retail, astronomy, and intelligence (Still et al., 2014).

2.2 *What information do managers consult?*

A growing body of evidence suggests that managers and executives seldom consult academic journals. Instead, they prefer to obtain knowledge of current trends in their respective fields mostly through marketing and business magazines and other 'grey literature' in the form of newsletters, fact sheets, pamphlets and conference abstracts produced mainly by practitioners (Bennett, 2007). In a survey of 141 marketing managers active in the UK computer industry, Bennett (2007) found that only 2% of the respondents read academic marketing journals and 3% reviewed marketing textbooks. Conversely, 89% of the interviewees read grey marketing literature while 62% read marketing magazines. At the same time, 58% of managers rated grey literature as either important or very important for identifying new ideas about marketing while 46% of the executives held similar views towards marketing magazines. Academic journals and textbooks, on the other hand, were deemed important by a mere 4% of the managers and considered very important by none (Bennett, 2007).

Similarly, Forster (2007) surveyed 87 business and public-sector leaders in Australia to determine whether these professionals consulted academic journals or practitioner-oriented magazines. In more than three-quarters of cases, respondents indicated that they had not even heard of many leading academic journals. By contrast, 67 respondents read *The Economist*, 65 read *Business Review Weekly*, 31 read *The Bulletin*, 24 read *Harvard Business Review* and 24 read *Financial Times* either 'sometimes' or 'regularly'. When asked why they read certain publications, 71 respondents indicated that they were "timely, topical, current and up to date with the industry or business sector" while 50 respondents reported that the selected publications were "concise, targeted, readable and had good executive summaries". When asked why they did not read certain publications, 43 participants stated that they contained "boring, dull, unreadable, turgid and inert language", 35 indicated that they were "too theoretical/impractical", 27 stated that they were "full of jargon", and 22 stated that they had "over-complicated statistics".

The sentiments expressed by Australian business and public-sector leaders are also reflected elsewhere. Stadler (2015) wrote in Forbes that business "leaders usually don't have the time to battle with the inaccessible prose of academic articles".

Scientists too are worried that academic papers are becoming so unintelligible that they have found some papers in their field too difficult to read. British biologist Richard Dawkins observed that "most papers in *Nature* and *Science* today can be read only by specialists in their respective fields" (Wyke, 2015). Harvard psychologist Stephen Pinker blames the 'professional narcissism' of many academics claiming that "I frequently find myself baffled by the writing in journals in my own specialty" (Wyke, 2015). Also, the University of Warwick data science fellow Adrian Letchford and his team of colleagues concluded that long titles and long-winded sentences often negatively impacted an article's influence and readership (Wyke, 2015).

So what publications do executives consult? Bill Gates, former chief of the Microsoft Corporation, for example, reads *The Wall Street Journal*, *The New York Times*, and *The Economist* cover to cover (Griswold and Nisen, 2014). Warren Buffett, a celebrated investor, reads *The Wall Street Journal* and *The New York Times* along with *USA Today*, *Financial Times*, and *American Banker* (Griswold and Nisen, 2014). Former US President Barack Obama reads *The Wall Street Journal*, *The New York Times*, and *The Washington Post* every morning and also enjoys *The New Yorker* and *The Atlantic* (Griswold and Nisen, 2014). Traditional academic and scientific literature seldom makes the reading list of leaders and decision makers.

Additional evidence comes from Leadtail, a social media marketing firm, which analysed 98,079 tweets in August 2015 from 1,713 executives comprising CEOs, CIOs, CFOs, and CMOs (Leadtail, 2015). The firm found that "C-Suite executives prioritise prominent business and news publications such as *Forbes*, *Business Insider*, *Fortune*, and *The Wall Street Journal*. They also turn to *TechCrunch*, *Mashable*, *Wired* and *Fast Company* to keep an eye on the latest technology and innovation stories". Again, there was scant evidence of, if any, executives tweeting about the latest publications from the academic press.

Although brief, yet the evidence presented above indicates that managers and executives stay updated on business and technological innovations by reviewing magazines and newspapers as opposed to academic papers and textbooks. This study did not collect direct evidence of what type of literature was consulted by the executives responsible for big data procurement. Instead, it infers from the reading habits of business executives in general.

3 Methods and data

The corpus text-mined in this study includes news stories published in the popular press during the eight years from 2010 to 2017.

3.1 Data collection

The news corpus was obtained from LexisNexis® Academic, which is the most widely-used news repository in social sciences (Weaver and Bimber, 2008). It covers over 26,000 news sources across the world in several languages (LexisNexis, n.d.). Table 2 documents different categories of news publications in LexisNexis, including sample titles.

Table 2 LexisNexis publication categories

Publication category	*Sample titles*
Newspapers	*The New York Times*, *The Washington Post*
News Transcripts	BBC, CNN
Newswires	The Associated Press, Xinhua
Magazines	*Forbes*, *Vanity Fair*
Newsletters	Hedgeweek, FDAnews publications
Web-based	efytimes.com, indiaretailnews.com
Journals	*Quarterly Journal of Economics*
Legal news	*Harvard Journal of Law and Public Policy*

The following query searched big data articles in LexisNexis:

```
Hlead(big data) AND Language(English) AND LENGTH(>600)
AND LENGTH(<2000) AND DATE (>2009/12/31) AND DATE
(<2018/01/01)
```

The above query returned news stories published in English that contained the term 'big data' in their title, highlight, or the lead paragraph; were 600 to 2000 words in length; and were published from the beginning of 2010 until the end of 2017.

The length of a news story is typically fewer than 2,000 words. For instance, *The Wall Street Journal* recommends the typical length of an op-ed article to be between 600 and 1,200 words (Wall Street Journal, 2017). Similarly, *The New York Times* recommends the ideal length of an op-ed to be 400 to 1,200 words (New York Times, 2017). A maximum length of 2,000 words filters out academic articles, which are substantially longer. Since LexisNexis also carries other material, such as short notices, which are often fewer than 600 words long, the choice of minimum length helps extract only the full-length articles.

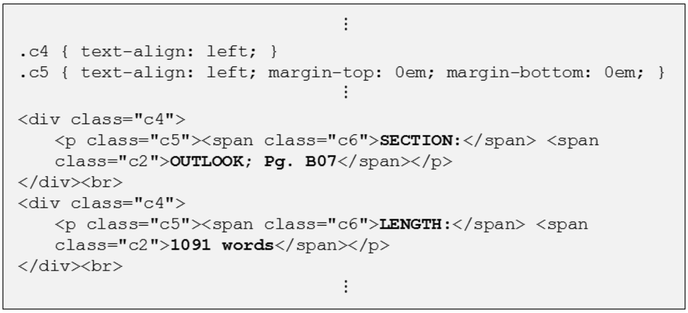
It was only in 2011 that big data became a popular term in the academic press (Gandomi and Haider, 2015). The same holds for news media such that our search of LexisNexis contents identified only eight news stories published in 2010. The number of big data stories grew exponentially starting in 2011.

3.2 Data preparation

In addition to the textual components of an article (i.e., its title and body), LexisNexis provides metadata for each story. Metadata refers to 'data about data' (Quemada and Simon, 2003), which in this case includes the publication date, type, and length. Metadata also contains subjects, industries, and locations discussed in the news article. LexisNexis automatically generates metadata mentioned above using a rule-based classification algorithm called SmartIndexing Technology (LexisNexis, 2013). A relevance score that varies between 50% to 99% signifies the extent of discourse that corresponds to each index for each article.

LexisNexis saves search results in HTML format with a maximum of 200 articles per file. HTML is a markup language primarily used to structure web pages. Figure 2 shows part of the source code of sample content. The first two lines in this figure are cascading style sheet (CSS) codes whereas HTML tags are enclosed in angle brackets.

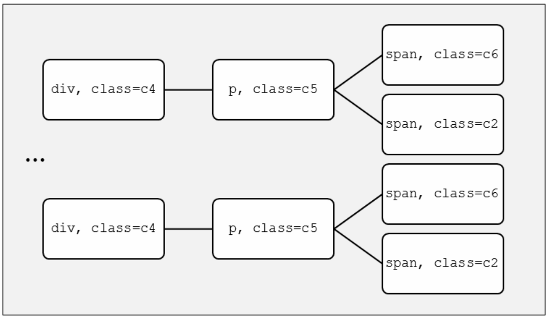
Figure 2 Partial source code of the search result



```
                                    ⋮
.c4 { text-align: left; }
.c5 { text-align: left; margin-top: 0em; margin-bottom: 0em; }
                                    ⋮
<div class="c4">
   <p class="c5"><span class="c6">SECTION:</span> <span
   class="c2">OUTLOOK; Pg. B07</span></p>
</div><br>
<div class="c4">
   <p class="c5"><span class="c6">LENGTH:</span> <span
   class="c2">1091 words</span></p>
</div><br>
                                    ⋮
```

HTML data are considered *semi-structured* (Abiteboul et al., 1997) because they lack the strict configuration of *structured* data (e.g., data stored in tables). However, the markup tags create a logical organisation in HTML data, which differentiates it from *unstructured* data (e.g., free text). To analyse the HTML data, semi-structured data must be converted into a structured format. To this end, the first step is to find the hierarchical organisation of data in the HTML file. This hierarchy can be represented in a parse tree, a tree with nodes labelled by HTML tags. Figure 3 displays the parse tree for the portion of the HTML code shown in Figure 2.

Figure 3 Parse tree of the HTML code shown in Figure 2



The next step is to navigate through the parse tree, capture the target information, and store it in a tabular format. This study used Python's Beautiful Soup library and regular expressions to parse the HTML files and commit the resulting content to tabular format. The textual and metadata for individual news items are thus captured and stored in four different tables comprising subjects, industries, geographical locations, and the remaining data.

3.3 Data analysis algorithms

This study uses text mining and NLP algorithms to analyse textual contents of the corpus. The content analysis involves two parts. First, the paper uses topic modelling to analyse the primary topics discussed in the news articles and analyse their trends over time. Second, this study develops an algorithm based on the distribution of *modal verbs* to quantify the tone and tenor in the coverage of big data.

3.3.1 Topic modelling

We used SAS for topic extractions. The process comprises four nodes. The 'file import' node converts the output from the data preparation stage into a SAS data source. The 'text parsing' node then extracts the structured information about the phrases comprising the corpus. Text parsing further entails three steps namely tokenisation, normalisation, and part-of-speech (POS) tagging. Tokenisation refers to parsing a string of characters into meaningful lexical elements, such as words, punctuation, numbers, and phrases (Hearst, 1997). A default dictionary of multi-word tokens supports the text-parsing node. Normalisation is the process of mapping tokens to their canonical form (e.g., mapping 'am, is, are, was, were, being, been' to 'be'). Normalisation reduces the size of the output from parsing (Chakraborty et al., 2013). The last step in text parsing determines the lexical category of each token (e.g., noun, verb, adverb) and labels them accordingly.

The text parsing node automatically detects the entity type and the token's attributes. The procedure also eliminates the terms that are on the 'stop list', which is a default set of low-information words, such as *be*, *say* and *not*.

Text parsing is followed by 'text filter', which computes a weighted term-by-document matrix that is an $n \times m$ matrix, where n is the number of all tokens in the corpus and m is the number of documents (i.e., the number of news stories). The term-by-document matrix facilitates the application of classical data mining techniques to analyse text data. For instance, one can deploy data mining methods to classify or cluster documents without considering their linguistic properties.

The element a_{ij} in the matrix represents the importance of the corresponding term in the corpus. A variety of weighting schemes have been reported in the literature (e.g., Hammouda and Kamel, 2004; Huang, 2008). Text filter's default weighting scheme, used in this study, is listed below:

$$a_{ij} = \log_2\left(f_{ij} + 1\right) \times w_i$$

where f_{ij} represents the frequency of term i in document j and w_i denotes the weight of term i, which is calculated based on the entropy measure shown below (Dumais, 1991):

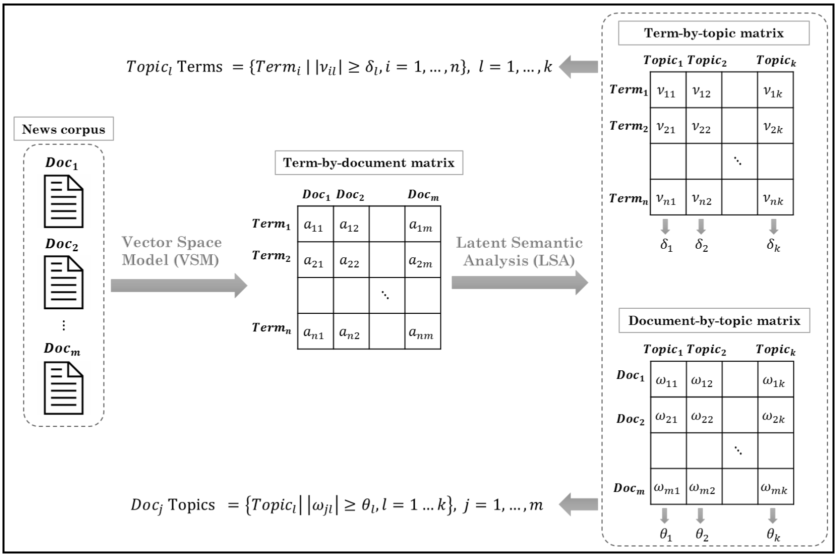
$$w_i = 1 + \sum_{j=1}^{n} \frac{\left(f_{ij}/g_i\right) \cdot \log_2\left(f_{ij}/g_i\right)}{\log_2 n}$$

In the above formula, g_i denotes the global frequency (within the corpus) of term i and n represents the number of documents in the corpus. This serves the purpose of assigning greater weights to the terms with a high local (within the document) and a low global frequency, because such terms have a greater discriminatory power.

The last node is the 'text topic' procedure. A topic is a collection of semantically related terms that characterise an idea or a theme. This node automatically extracts topics and uses a degree of association to assign terms and documents to topics. Topic extraction is an example of soft clustering, since a document may be associated with multiple topics or none at all, as opposed to the hard clustering where each document is assigned to exactly one cluster (Aliguliyev, 2009).

Figure 4 depicts topic extraction in the SAS text topic node. The theoretical basis of the method is detailed in Appendix 1. Like many other text mining techniques, it employs the vector space model (VSM), which represents documents as vectors in the term space (Salton et al., 1975). Specifically, in a corpus with n terms, each document is expressed as an n-dimensional vector whose coordinates are the weights of the corresponding terms in the document. Thus, vectors are same as the columns of the weighted term-by-documents matrix.

Figure 4 Topic extraction procedure


$Topic_l$ Terms $= \{Term_i \mid |v_{il}| \geq \delta_l, i = 1, \ldots, n\},\ l = 1, \ldots, k$
News corpus
Doc_1
Doc_2
Doc_m
Vector Space Model (VSM)
Term-by-document matrix
Doc_1 Doc_2 Doc_m
$Term_1$ a_{11} a_{12} a_{1m}
$Term_2$ a_{21} a_{22} a_{2m}
$Term_n$ a_{n1} a_{n2} a_{nm}
Latent Semantic Analysis (LSA)
Term-by-topic matrix
$Topic_1$ $Topic_2$ $Topic_k$
$Term_1$ v_{11} v_{12} v_{1k}
$Term_2$ v_{21} v_{22} v_{2k}
$Term_n$ v_{n1} v_{n2} v_{nk}
δ_1 δ_2 δ_k
Document-by-topic matrix
$Topic_1$ $Topic_2$ $Topic_k$
Doc_1 ω_{11} ω_{12} ω_{1k}
Doc_2 ω_{21} ω_{22} ω_{2k}
Doc_m ω_{m1} ω_{m2} ω_{mk}
θ_1 θ_2 θ_k
Doc_j Topics $= \{Topic_l \mid |\omega_{jl}| \geq \theta_l, l = 1 \ldots k\},\ j = 1, \ldots, m$

In large corpora, VSM vectors are high-dimensional and sparse because many terms occur only in a few documents. As a result, processing the corpus in its full dimensional space often proves computationally challenging. Researchers have therefore developed several dimensionality reduction techniques. Latent semantic analysis (LSA) is a popular technique, which is also deployed by the SAS text topic node (Deerwester et al., 1990).

LSA applies singular value decomposition (SVD), a linear algebraic method for matrix factorisation, to decompose the sparse term-by-document matrix into dense factors (Mobasher et al., 2004). LSA keeps the first k dimensions of the SVD components and eliminates the rest. The basic idea is to map the data from the original space to a lower dimension while maintaining its important features. In this abstract space, each SVD dimension represents a latent topic in the corpus. As illustrated in Figure 4, SVD generates two factors from the text-by-document matrix, namely a term-by-topic matrix and a document-by-topic matrix.

As explained in Appendix 1, for each term in the corpus, the algorithm assigns an association weight concerning each topic. The term cutoffs are then used to determine the collection of the terms describing each topic. Similarly, the algorithm uses document-topic association weights and the corresponding document cutoff thresholds to specify the topics covered in each document.

3.3.2 Tone and tenor analysis

Modal verbs are semantic/grammatical features of the language used to express modality (e.g., *can*, *could*, *may*, *might*, *must*, *will* and *would*) (Sanz, 2011). Modal verbs convey the writer/speaker's opinion or judgment toward the future (Neff et al., 2003). According to Biber et al. (1999), as cited in Dafouz et al. (2007), modal verbs, based on their semantic meaning, fall into three groups: necessity/obligation (e.g., *must*, *should*), ability/likelihood/permission (e.g., *can*, *could*, *may*, *might*), and volition/prediction (e.g., *will*, *would*).

Thus, one can semantically analyse the tone of big data stories reported in the press by analysing the overall frequency of modal verbs in the corpus and their frequency in sentences where 'big data' is the subject. This section documents the procedures devised to extract modal verbs from 'big data' sentences. The next section details the method developed for the analysis.

The following algorithm outlines the procedure for extracting modal verbs from sentences 'big data' is the subject.

Algorithm 1 Extracting modal verbs associated with 'big data'

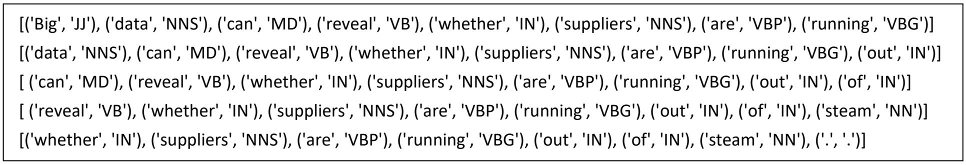
```
0:  big_data_modal_varbs = [ ]
1:  for each story in the corpus do:
2:  Segment the story into sentences;
3:      for each sentence do:
4:          if 'big data' and a modal_verb are in the sentence then:
5:              Tokenise the sentence into words;
6:              Find the POS tag for each word;
7:              Produce all the n-grams in the sentence;
8:              for each n-gram do:
9:                  if the n-gram begins with 'big data' then:
10:                     Parse the n-gram with the chunk grammar;
11:                     Extract the new_modal_verb from the parse tree;
12:                     Add the new_modal_verb to big_data_modal_varbs;
13:                 end if
14:             end for
15:         end if
16:     end for
17: end for
```

In steps 2, 5, and 6, the study used NLTK's (Natural Language Toolkit) sentence tokeniser, word tokeniser, and part-of-speech tagger, respectively. The POS tagger uses the Penn TreeBank tagset, which includes 36 tags (e.g., 'MD' for modal, 'RB' for an adverb, and 'JJ' for adjective).

Steps 8 to 10 in the algorithm produce all possible *n*-grams in the sentence and retains those starting with 'big data'. With a small *n*, one may miss the modal verb. Higher *n* values, on the hand, return an empty set for the sentences shorter than n tokens. To find the optimal *n*, we manually extracted the relevant modal verbs in a small sample of the corpus. We then ran the algorithm for all *n* values in the range [4, 10]. We found *n* = 8 yielded the best match to the manual coding with an error of zero. Thus, we ran the code with 8-grams in step 8. The algorithm is illustrated with an example below.

Consider the sentence 'big data can reveal whether suppliers are running out of steam'. Step 7 generates five 8-grams for this sentence, which are listed in Figure 5 along with the POS tags. The first 8-gram starts with 'big data' and therefore are fed to steps 10–12.

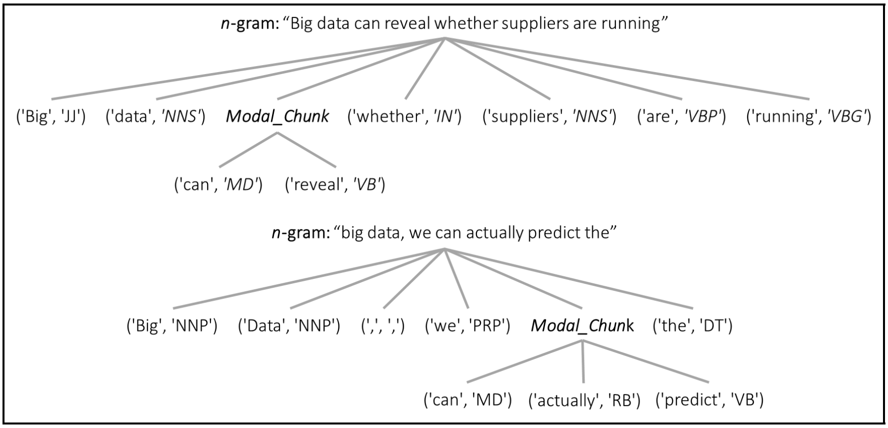
Figure 5 8-grams in the example sentence



```
[('Big', 'JJ'), ('data', 'NNS'), ('can', 'MD'), ('reveal', 'VB'), ('whether', 'IN'), ('suppliers', 'NNS'), ('are', 'VBP'), ('running', 'VBG')]
[('data', 'NNS'), ('can', 'MD'), ('reveal', 'VB'), ('whether', 'IN'), ('suppliers', 'NNS'), ('are', 'VBP'), ('running', 'VBG'), ('out', 'IN')]
[ ('can', 'MD'), ('reveal', 'VB'), ('whether', 'IN'), ('suppliers', 'NNS'), ('are', 'VBP'), ('running', 'VBG'), ('out', 'IN'), ('of', 'IN')]
[ ('reveal', 'VB'), ('whether', 'IN'), ('suppliers', 'NNS'), ('are', 'VBP'), ('running', 'VBG'), ('out', 'IN'), ('of', 'IN'), ('steam', 'NN')]
[('whether', 'IN'), ('suppliers', 'NNS'), ('are', 'VBP'), ('running', 'VBG'), ('out', 'IN'), ('of', 'IN'), ('steam', 'NN'), ('.', '.')]
```

Regular-expression chunking (see, e.g., Bird, 2006) extracts modal verbs from the 'big data' *n*-grams. The method requires a grammar rule (called chunk grammar) that delimits the target chunk in the *n*-gram. A chunk is a subset of the tokens that satisfy the chunk grammar, which in this case is 'a modal verb followed by zero or more adverbs and then one or more verbs'. Adverbs are included in the chunk grammar since they may occur between the modal and the main verb. For instance, consider the sentence "With big data, we can *actually* predict the thing they will need most" (Tutty, 2015). The chunker parses the *n*-gram with the given chunk grammar and builds a parse tree. Figure 6 illustrates the parse tree for the two example sentences presented above. Finally, in steps 11 and 12, the algorithm traverses the parse tree to find and store the modal verb.

Figure 6 Parsing tree for the example sentences



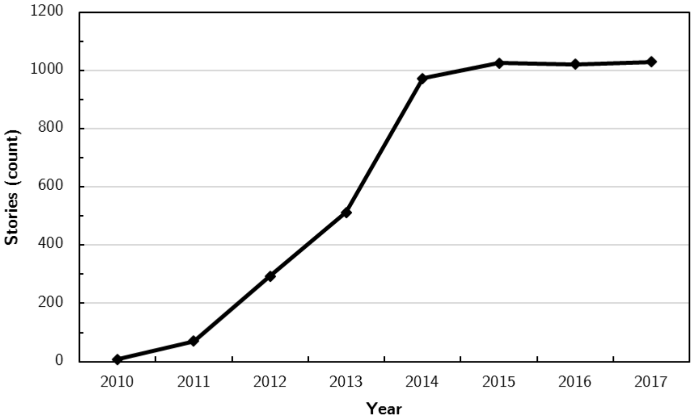
4 Results

Results are presented in two sections. The analysis of metadata, generated from LexisNexis's proprietary algorithms is presented first. The second part comprises the content analysis, which is further disaggregated into two subsections.

4.1 Results from metadata

The corpus comprised of 4,931 news items reported in major English news outlets and contained the term 'big data' in their title, highlight, or the lead paragraph. Starting in 2010 with just eight stories, big data coverage peaked in 2017 with 1,029 news items. The coverage of big data stabilised in 2015 such that the number of stories increased by only four in the following two years. A quick look at Figure 7 suggests an S-curve, indicative of a maturing market for big data.

Figure 7 News coverage of big data suggests a mature market with an S-curve



The average length of a news item was 925 words with the maximum length of 2,000 words. Almost 69% of the material came from newspapers. A little over 12% was sourced from web publications and around 8% from press releases carried by Newswire.

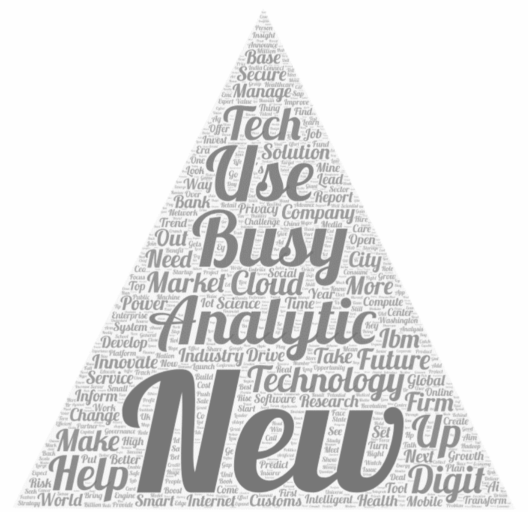
4.1.1 What do the headlines proclaim?

A typical reader is attracted to a news item by its headline. A carefully crafted headline is rewarded with the readers' attention. In journalism, the editors and not the writers craft the headlines. Therefore, an analysis of headlines carries additional value because it reflects the mindset of editors and not necessarily that of writers.

Figure 8 presents the most frequently mentioned attributes of big data in headlines in a word cloud such that the size of each attribute is a function of how frequently it was mentioned in the headlines. The most frequently mentioned terms include *new*, *analytic*, *busy*, *cloud*, *technology*. The single most prominent attribute in headlines (excluding the term 'big data') was *new*. Figure 8 reveals that the novelty of technology is what the editors highlighted in the headlines. At the same time, the editors relied on the technological attributes of big data as is reflected in the terms *analytic*, *cloud* (storage) and *technology* to attract readers.

Figure 8 also provides the first clue to the questions being explored in this study. What attracted readers to big data are the headlines projecting the newness of the big data technologies. Other headlined attributes included the capacity to store large amounts of data (cloud storage) and the ability to analyse data and draw inferences and insights.

Figure 8 Word cloud of text in the headlines of big data stories



Source: http://www.wordart.com

4.1.2 Primary subject areas in big data stories

LexisNexis uses proprietary algorithms to generate scores for various predefined subject areas that are most relevant to each news item. Table 3 presents the most relevant subject areas based on the relevance score.

Since 'big data' was part of the search term, the most pertinent subject identified by the algorithm ended up being big data. This finding was of course expected and hence is not very revealing. However, the remaining subjects offer valuable insights. The second most relevant subject is 'executives' suggesting that big data news coverage focused on corporate decision-makers. The underlying hypothesis in this study is that decision makers and business executives consulted news media about the virtues of big data and analytics and that the news reporting of big data was therefore targeted at the business executives. Despite 'executives' not being part of the search term used to extract the corpus, still, it emerged as the second most relevant subject in big data coverage. This finding offers support for the research hypothesis that big data news coverage was targeted at business executives.

The third most frequent subject was 'computer software'. Since the discourse on big data has focused on data storage, advanced search algorithms, and computing platforms, computer software was expected to rank among the top. Other relevant subject areas included 'business analytics', 'cloud computing' and 'artificial intelligence'.

Table 3 Subject index

Subject	*Count of stories*
Big data	3,364
Executives	1,425
Computer software	1,305
Business analytics	1,278
Cloud computing	777
Internet and WWW	734
Artificial intelligence	697
Computer networks	667
Computing and information technology	654
Social media	640
Mobile and cellular telephones	617
Internet social networking	599
Banking and finance	574
Information security and privacy	545
Internet of things	536

4.1.3 Big data industrial sectors

Table 4 highlights the sectors most relevant to big data using LexisNexis' proprietary algorithms. Surprisingly, the *information* sector emerged as the second most relevant industrial sector. The apriori was to expect the information sector at the top because the businesses focused on big data and analytics are essentially involved in collecting, storing, and analysing information. Another surprise was *manufacturing*, the top-ranked sector, which ranked higher than services. The widespread big data news coverage is largely known for the coverage of consumer behaviour, retail (fifth most relevant), and finance/insurance (fourth most relevant). Yet, *Manufacturing* emerged as the most frequently cited sector in the corpus perhaps because big data hardware involves manufacturing innovative and scalable solutions to store the rapidly growing digital footprint.

Also note that the *services* sector does not appear as a standalone economic sector in Table 4, as it is proxied as *professional*, *scientific* and *technical services*, which emerges as the third most frequently cited economic sector. One can argue that *retail*, *finance* and *insurance*, *administration services* and other similar sectors also fall under the general rubric of the *services* sector, which would make *services*, when taken together, even more pronounced, if not the top-cited, sector in Table 4.

Table 4 Most frequently cited economic sectors (NAICS 2-digit classification)

Industrial sectors	*Count of stories*
Manufacturing	2,277
Information	2,037
Professional, scientific, and technical services	1,184
Finance and insurance	831
Retail trade	303
Administrative and support and waste management and remediation services	128
Wholesale trade	120
Real estate rental and leasing	115
Transportation and warehousing	100
Mining	100
Management of companies and enterprises	97
Utilities	48
Accommodation and food services	31
Construction	29
Arts, entertainment, and recreation	16
Educational services	10

Table 5 Big data geography

Cities	*Country*	*Count of stories*
San Francisco	USA	453
Silicon Valley	USA	220
New York	USA	150
Singapore	Singapore	130
Beijing	China	121
London	UK	120
New Delhi	India	106
Bangalore	India	83
Shanghai	China	75
Boston	US	60
Toronto	Canada	55
Mumbai	India	53
Sydney	Australia	45
Dublin	Ireland	33

4.1.4 The geography of big data

Big data news coverage highlights the unique geography reflective of the intellectual capital associated with research and development in big data. A search for most frequently mentioned geographies related to big data revealed the usual suspects

(Table 5). The two most commonly mentioned geographies in the corpus were San Francisco and the Silicon Valley followed by New York such that the top three locations were based in the USA. Interestingly, Singapore, being a nation-state of fewer than 10 million and a land mass of just 280 square miles, appears fourth on the list, ahead of Chinese and European centres of excellence. London tops the list from Europe and Beijing from China. At the nation-state level, the USA gets the most mentions among the top 14 referenced geographies in the corpus followed by India and China.

4.2 Content analysis

4.2.1 Big data: from terms to topics

We applied the topic extraction algorithm explained earlier to the corpus of news stories covering the period 2010–2017. Appendix 2 and Table 6 present the results. Appendix 2 lists the top 10 descriptive terms for each topic. We subjectively assigned a title to each combination of terms to convey their collective meaning. Table 6 orders the topics based on how frequently they appeared in the corpus.

Table 6 Topics extracted from the entire corpus of stories in the period 2010–2017

Topic	*Overall frequency*
Evidence-based decision making	799
Big data start-ups	783
Digital economy	779
Retail analytics	682
Big data research	666
Smart cities	646
Cloud computing	632
Big data ROI	619
Privacy and security	607
Big data education	509
China’s investment on big data	441
Big data in lending industry	422
Healthcare analytics	386
Analytics job market	379
Smart farming	299
Electoral analytics	228

The most frequently cited topic in the corpus is *evidence-based decision making*. The topic is derived from the terms *analytics*, *insight*, *big data*, *analysis*, *unstructured*, *tool*, *organisation*, *predictive*, and *decision*, among others, which proxy the practice of data-driven decision making in organisations. A natural evolution for big data was to graduate from being just raw source to insights and finally leading to execution. Thus, evidence-based decision-making emerging as the leading topic is indicative of how the news media saw and promoted the intended use of big data technologies. The second topic relates to *big data start-ups*, reflecting the rise of venture capital invested in analytics-oriented start-ups across industrial sectors.

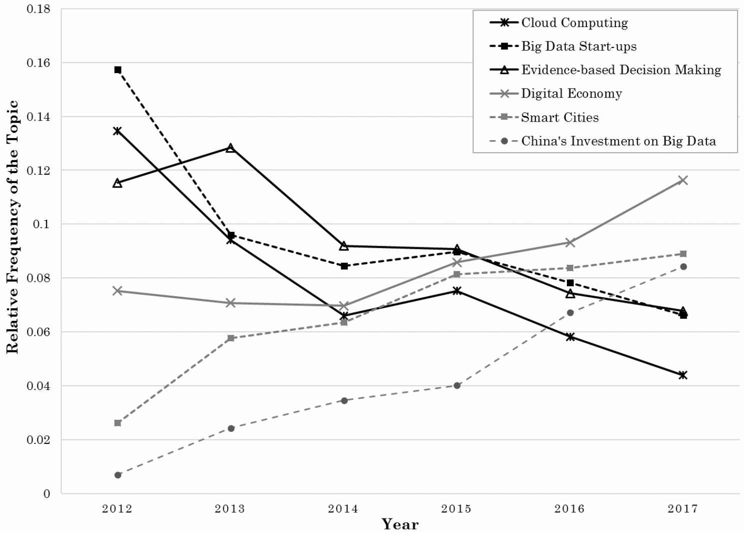
A common thread running through the big data stories is the potential for customer analytics to increase sales and profits. Hence, *retail analytics* emerges as the fourth most-frequently reported topic. At the same time, concerns about privacy, data security, and the legal frameworks related to big data also appeared in the top 10 most frequently reported topics in the corpus. A case in point is the recent backlash against Alphabet's (Google's parent company) plans to develop part of downtown Toronto as a smart city prototype. A lack of transparency in data collection, storage, use, and governance was the subject of the mainstream coverage of big data in Canada (Canon, 2018).

Together, *big data education* and *analytics job market* capture the initial and continued concerns about the shortage of experts needed to sustain a data-driven corporate culture. Industry surveys forecasted millions of unfilled positions requiring big data expertise in the USA alone (Manyika et al., 2011). The news media, as mentioned earlier, had bought into the scarcity of talent narrative that initially appeared as a widely reported publication by the McKinsey Global Institute.

The other key themes that emerged in the corpus included *digital economy*, *smart cities*, *cloud computing*, *healthcare analytics*, *electoral analytics* and *smart farming*.

This study also analyses the temporal trends in the relative frequency of topics during the 2012–2017 period. The procedure to determine trends involved two steps: first the relative frequency of topic i in year j was obtained by $f_i \Big/ \sum_i f_{ij}$, where f_{ij} is the raw frequency of the topic i in year j. Then a least-square regression line was fit on each topic with year as the explanatory variable and relative frequency as the dependent variable. The slope of the line represents the direction and magnitude of the trend. Figure 9 shows the trend for the top three increasing and declining topics over time. The reporting about *big data start-ups*, *evidence-based decision making*, and *cloud computing* has been gradually declining while the coverage of China's investment in big data, smart cities, and digital economy have been on the rise.

Figure 9 Major topic trends in the coverage of big data



What is evident from the preceding observation is that the big data coverage in the news media has not been static, nor has it remained focussed on the initial themes about the promise of big data and the scarcity of human capital. Instead, it has evolved further to embrace newer challenges and recognises new players in the big data domain. One can see evidence that the news coverage has transformed as it highlights big data applications in smart cities and the shifting nucleus of big data R&D and application from the western countries to East Asia.

4.2.2 The tone of big data stories

The purpose behind the content analysis of the 5-million-word corpus is to determine how the news media presented information to the readers. As discussed earlier, the journalistic account of big data, owing to its newness and claims about the great promise by the proponents, could be promotional rather than being critical. Hence, instead of looking for and reporting on the proof of concept, the reporting could echo the claims made by the big data proponents. To explore these concerns, one would have to analyse the modal verbs to determine the tone and tenor of big data coverage.

This study extracts modal verbs to determine how the news coverage highlights the promise of the new technology, e.g., big data 'can' or 'will'. Also, of interest is the change over time in tone as big data became popular over the years. The analysis presented here is further differentiated as it distinguishes between the tone in the entire corpus and the tone expressed in sentences that include the term 'big data'.

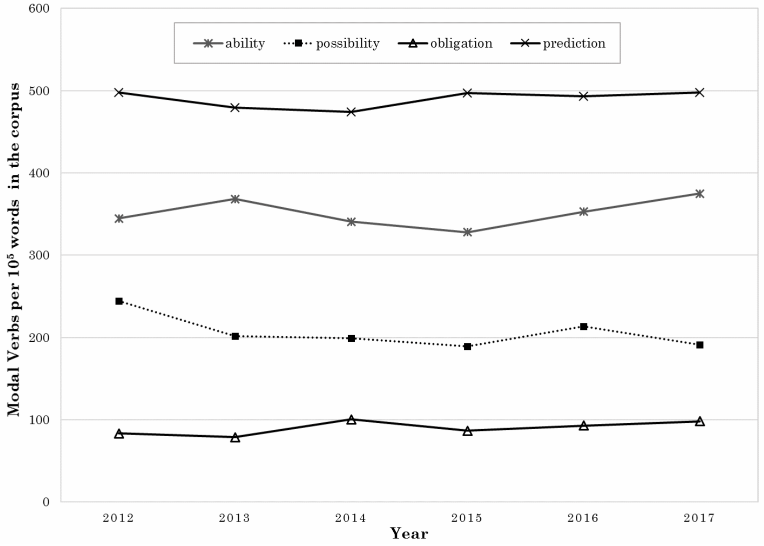
This study aggregates certain modal verbs to present specific trends over time. The goal is to determine how the narrative evolved. Some modal verbs, such as *could*, *may*, and *might* represent the possibilities resulting from deployment of big data. One can see how frequently such modal verbs appeared each year relative to all words comprising the corpus for that year. Similarly, one can determine how frequently the modal verbs representing possibilities appeared in sentences mentioning big data relative to all words comprising the corpus in a year.

Similarly, other modal verbs, such as *shall*, *will*, and *would* are indicative of big data's promise for future. The relative frequency of predictive modal verbs and their change over time in individual stories or specific sentences mentioning the word big data would shed light on whether the content of stories changed from predictions about big data-enabled outcomes to perhaps proof-of-concept reporting about what big data enabled.

At the same time, other modal verbs, such as *must* and *should* indicate what has been expected of big data. For instance, 'big data must' accomplish some goal is indicative of an obligation or necessity. The relative frequency of modal verbs representing obligation will again shed light on how the coverage changed over time. Lastly, the ability of big data is captured by the modal verb *can*.

Figure 10 presents the relative frequency of modal verb trends over time in big data stories (the entire corpus) for ability, obligation, prediction, and possibility after being normalised by the total number of words comprising the corpus for that year. For instance, the modal verb depicting ability (i.e., *can*) appeared 344.56 times in the entire corpus for every 100,000 words in 2012. This is obtained by dividing the frequency of *can* in 2012 stories (1,056) by the entire corpus word count (in 100,000s) in that year (3.06482). Figure 10 reveals that the modal verbs for ability, possibility, obligation, and prediction depict a flat trend from 2012 to 2017. The test of significance for the slope of a least-squares regression line yields a p-value of 0.99, providing strong evidence against the existence of a trend. This would lead one to conclude that the tone of the news coverage did not change over time.

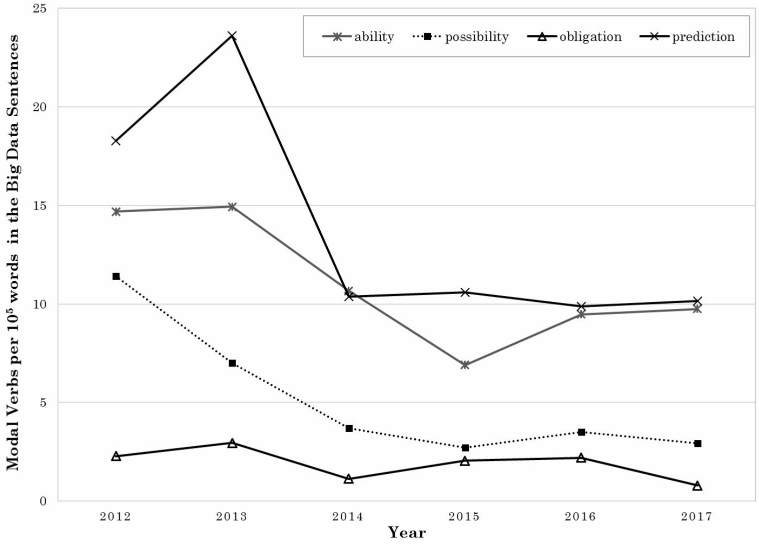
Figure 10 Model verbs trends over time in big data stories



In comparison, a revealing picture emerges in Figure 11 that presents the modal verb trends in sentences that contained the phrase 'big data' as the subject (refer to Section 3.3.2 for the extraction procedure). For instance, the model verb 'can', which depicts ability, appeared 14.94 times in big data sentences for every 100,000 words in the entire corpus in 2013. The figure suggests the presence of a trend in the tone of big data sentences. The test of slope significance yields a p-value of 0.026 offering evidence of the existence of temporal trends in news coverage. One can see that model verbs representing prediction (shall, will) increased in frequency from 2012 to 2013, but declined in 2014 and have stayed at the same level since then. This is indicative that the initial big data news coverage was relatively more predictive highlighting the virtues of the new technology and the promise it held for innovation. One can infer a similar declining trend for modal verbs depicting *possibility* from a higher level in 2012 to a lower level in 2015. The modal verbs serving as proxies for obligation have stayed the same in their relative frequency in big data sentences over time.

The analysis presented here demonstrates that if one were to analyse the modal verbs at the entire corpus level, one would miss the trends latent in sentences that contained the expression big data. This study uses an innovative approach by analysing the expressions of interest at the sentence level rather than at the document level. The current practice of content analysis processes terms of interest at the document level that might miss the context in which the terms have been used. The modal verb analysis presented here thus reveals that when one focuses on the sentences that mentioned big data, one could see that the tone of big data coverage changed over time with the earlier focus on the possible virtues and abilities of big data declining over time.

Figure 11 Model verbs trends over time in big data sentences



5 Conclusions

Over the past few years, businesses have spent tens of billions of dollars in acquiring hardware and software solutions to harness insights from data they have either generated or collected. Such massive spent on big data technologies, which took off in 2011, was realised in a short span of a few years.

This study reviewed the information sources regularly consulted by business leaders and executives. It further explored the content of the news media to determine the tone and tenor of the information available to big data executives.

Despite time constraints because of their busy schedules, managers and executives must acquire new knowledge to inform future decision making. Academic and social media research suggests that business leaders rarely consult scholarly publications to inform themselves. The literature reviewed in this study about the reading habits of senior executives and managers revealed that they preferred reviewing news media and professional literature over scholarly publications.

A review of the academic press citations and news media during 2010 and 2017 revealed that academic publications were slow in producing timely reviews of big data theories and applications. News media and industry publications filled that void. Whereas the academic press has been the key source for trusted and unbiased advice on computing technologies, large databases, and scalable analytics, the popular press became the primary channel for spreading awareness about ‘big data’ as a marketing concept. At times, the news media helped popularise the innovative ideas published in the academic press.

Realising that executives rely more on news media than scholarly publications for information and that the scholarly press was busy playing catch-up on big data coverage, the content analysis of news media's coverage of big data therefore helps understand what informed the decision making of business leaders.

This study found that big data news coverage focused on the newness of the emerging paradigm and highlighted its technological aspects. The most frequent attributes mentioned in the 4,931 headlines included *new*, *analytic*, *busy*, *tech/technology*, and *cloud*. The content analysis revealed that executives were the targeted readership of the big data coverage, which was evidenced by 'executives' returning the second-highest relevance score after the term big data. Computer software, business analytics, and cloud computing also returned high relevance scores for subjects covered in big data stories. Similarly, the news coverage zeroed in on a select few industries, namely *information*; *professional*, *scientific*, and *technical services*; and *manufacturing*.

The analysis of the geography of big data confirms the informally held views about certain tech centres. As expected, San Francisco and Silicon Valley were the most frequently mentioned geographies in the corpus followed by New York. Outside of North America, Singapore, London, and Beijing were frequently mentioned.

The text mining algorithms extracted the most relevant topics from the corpus. The topics centred on evidence-based decision making, start-ups, digital economy, cloud computing, healthcare analytics, electoral analytics, and smart farming/agriculture, with topics such as the digital economy, smart cities, and China's investment on big data on the rise.

Furthermore, the study found that news coverage was more about the promise of big data than the proof of concept. The analysis of modal verbs revealed that the tone of big data coverage in the beginning focussed on the predictions about what big data would achieve in the future. However, the big data discourse predicting the future abilities and possibilities of big data has declined over time.

A critical review of content analysis presented here has identified big data clichés, such as projecting big data as a solution for almost all that ails the economy and society. News stories were consistent in singing the praise of the same technologies, geographies, and processes. The coverage seems repetitive and uncritical. This is not to suggest that big data-driven analytics hold no or limited value. The concern is about the uncritical trust in big data's 'unlimited' potential to discover cures for illnesses or solutions for socio-economic challenges, such as economic inequality or traffic congestion.

The news media reported verbatim estimates from the industry generated reports. The management consulting firm McKinsey and Company had no fewer than 200 mentions in the corpus with numerous news reports repeating McKinsey's estimate of an impending shortage of 140,000 to 190,000 analytic experts and 1.5 million data fluent business managers in the USA. Again, news reports cited McKinsey for the claim that "companies that put data at the centre of the sales and marketing decisions improved their marketing ROI by 15 to 20 percent" or another news item citing McKinsey for "businesses that embrace digital transformations could see increased revenue as much as 30 percent". McKinsey was again the source for the claim that "open data could add more than $3 trillion in total value annually".

At a time when the executives were searching for evidence, it appears big data news coverage was mostly about the usual suspects and was of promotional nature than being critical and reflective. The conformity in coverage, as is evidenced by the repetitive mention of the same geographies and industries, is likely to have influenced the big data procurement decisions.

No wonder, recent attempts to determine the return on investment (ROI) of big data procurement, as discovered in the review of the literature, have returned mixed results with most firms either not been able to determine the ROI or obtained less than ideal returns. For instance, Davenport and Dyché (2013) found only a small number of firms maintained meticulous records for their ROI on big data analytics. Shim et al. (2015) also found that in their haste to implement big data solutions some businesses ignored setting up metrics to determine the ROI. Research has shown that 70% of the big data implementations are deemed unsuccessful because managers ignored implementing appropriate metrics to determine ROI and, at times, failed to gain executive support for subsequent expansions (Bertolucci, 2013).

Similarly, the 2017 Annual Survey of Big Data Business Impact by NewVantage Partners, a management consulting firm focused on business innovation, revealed that whereas 95% of the business executives reported that their firms had undertaken a big data project, only 48.4% reported achieving measurable results from their big data investments (NewVantage, 2017). Most respondents reported that the results from big data investments were moderately successful, too early to tell, or failure.

This study has made the following contributions. It conducted a systematic review of the contents of a large corpus of nearly five million words to identify the themes, tone, and tenor of the news coverage. The study identified the industries and geographies referenced the most in big data news coverage. Lastly, and more importantly, the study has described in some detail the theory and application of text mining and NLP algorithms.

The novel application of statistical methods (i.e., regression models to determine temporal trends) to the output of NLP algorithms allows one to determine how topics extracted from the corpus changed over time. Furthermore, this study uses an innovative approach in NLP by analysing the expressions of interest at the sentence level rather than at the document level. The analysis reveals that when one focuses on the sentences that mentioned big data, one could see that the tone of big data coverage changed over time with the earlier focus on the possible virtues and abilities of big data declining over time. The methodological details and their applications serve as a template for other researchers who will value an illustrated use of advanced text mining tools.

References

Abiteboul, S., Quass, D., McHugh, J., Widom, J. and Wiener, J.L. (1997) 'The Lorel query language for semistructured data', *International Journal on Digital Libraries*, Vol. 1, No. 1, pp.68–88.

Aliguliyev, R.M. (2009) 'Performance evaluation of density-based clustering methods', *Information Sciences*, Vol. 179, No. 20, pp.3583–3602.

Agarwal, N. and Brem, A. (2015) 'Strategic business transformation through technology convergence: implications from General Electric's industrial internet initiative', *International Journal of Technology Management*, Vol. 67, Nos. 2–4, pp.196–214.

Begdache, L., Kianmehr, H. and Heaney, C.V. (2018a) 'College education on dietary supplements may promote responsible use in young adults', *Journal of Dietary Supplements*, pp.1–14.

Begdache, L., Kianmehr, H., Sabounchi, N., Chaar, M. and Marhaba, J. (2018b) 'Principal component analysis identifies differential gender-specific dietary patterns that may be linked to mental distress in human adults', *Nutritional Neuroscience*, pp.1–14.

Bennett, R. (2007) 'Sources and use of marketing information by marketing managers', *Journal of Documentation*, Vol. 63, No. 5, pp.702–726.

Bertolucci, J. (2013) 'Big data ROI still tough to measure', *InformationWeek* [online] http://www.informationweek.com/big-data/news/big-data-analytics/big-data-roi-still-tough-to-measure/240155705 (accessed 2 December 2016).

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*, Vol. 2, Longman, Harlow, UK.

Bird, S. (2006) 'NLTK: the natural language toolkit', in *Proceedings of the COLING/ACL on Interactive Presentation Sessions, COLING-ACL '06*, pp.69–72, Association for Computational Linguistics.

Canon, G. (2018) ''City of surveillance': privacy expert quits Toronto's smart-city project', *The Guardian*, 23 October [online] http://www.theguardian.com/world/ 2018/oct/23/toronto-smart-city-surveillance-ann-cavoukian-resigns-privacy (accessed 25 December 2018).

Chakraborty, G., Pagolu, M. and Garla, S. (2013) *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*, SAS Institute, Cary, NC.

Columbus, L. (2015) 'Roundup of analytics, big data & business intelligence forecasts and market estimates, 2015', *Forbes* [online] http://www.forbes.com/sites/louiscolumbus/2015/05/25/roundup-of-analytics-big-data-business-intelligence-forecasts-and-market-estimates-2015/ (accessed 25 November 2017).

Dafouz, E., Nez, B. and Sancho, C. (2007) 'Analysing stance in a CLIL university context: non-native speaker use of personal pronouns and modal verbs', *International Journal of Bilingual Education and Bilingualism*, Vol. 10, No. 5, pp.647–662.

Davenport, T.H. and Dyché, J. (2013) *Big Data in Big Companies* [online] http://docs.media.bitpipe.com/io_10x/io_102267/item_725049/Big-Data-in-Big-Companies.pdf (accessed 2 December 2016).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. (1990) 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp.391–407.

Diebold, F.X. (2012) *A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version (November 26, 2012)*, PIER Working Paper No. 13-003 [online] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843 (accessed 17 November 2017).

Dumais, S.T. (1991) 'Improving the retrieval of information from external sources', *Behavior Research Methods, Instruments, & Computers*, Vol. 23, No. 2, pp.229–236.

Fan, W. and Bifet, A. (2013) 'Mining big data: current status, and forecast to the future', *ACM SIGKDD Explorations Newsletter*, Vol. 14, No. 2, pp.1–5.

Forster, N. (2007) 'CEOs' readership of business and management journals in Australia: implications for research and teaching', *Journal of Management & Organization*, Vol. 13, No. 1, pp.24–40.

Gandomi, A. and Haider, M. (2015) 'Beyond the hype: big data concepts, methods, and analytics', *International Journal of Information Management*, Vol. 35, No. 2, pp.137–144.

Griswold, A. and Nisen, M. (2014) *What 16 Successful People Read in the Morning* [online] http://www.businessinsider.com/successful-people-morning-reading-habits-2014-1 (accessed 17 November 2016).

Hammouda, K.M. and Kamel, M.S. (2004) 'Efficient phrase-based document indexing for web document clustering', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, No. 10, pp.1279–1296.

Hearst, M.A. (1997) 'TextTiling: segmenting text into multi-paragraph subtopic passages', *Computational Linguistics*, Vol. 23, No. 1, pp.33–64.

Huang, A. (2008) 'Similarity measures for text document clustering', *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, Christchurch, New Zealand, pp.49–56.

IDC (2015) *New IDC Forecast Sees Worldwide Big Data Technology and Services Market Growing to $48.6 Billion in 2019, Driven by Wide Adoption across Industries* [online] http://www.idc.com/getdoc.jsp?containerId=prUS40560115 (accessed 25 November 2017).

Janssen, M., van der Voort, H. and Wahyudi, A. (2017) 'Factors influencing big data decision-making quality', *Journal of Business Research*, Vol. 70, pp.338–345.

Kelly, J. (2015) *Big Data Market Size and Vendor Revenues* [online] http://wikibon.org/wiki/v/Big_Data_Market_Size_and_Vendor_Revenues (accessed 25 November 2017).

Kristal, M. (2012) 'Mining mountains of data is key for Canadian businesses', *The Globe and Mail* [online] http://www.theglobeandmail.com/report-on-business/economy/canada-competes/mining-mountains-of-data-is-key-for-canadian-businesses/article4540604/ (accessed 25 November 2017).

Leadtail (2015) *What C-Suite Execs Read Every Day* [online] http://newvantage.com/wp-content/uploads/2017/01/Big-Data-Executive-Survey-2017-Executive-Summary.pdf (accessed 30 May 2017).

LexisNexis (2013) *Introduction to LexisNexis SmartIndexing TechnologyTM* [online] https://www.lexisnexis.com/infopro/resource-centers/product_resource_centers/b/smartindexing/archive/2013/09/11/what-is-lexisnexis-smartindexing-technology.aspx (accessed 13 April 2016).

LexisNexis (n.d.) *Nexis.com Comprehensive Online News & Business Research Tool* [online] https://www.lexisnexis.com/en-us/products/nexis.page (accessed 7 April 2016).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Technical report, McKinsey Global Institute [online] http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (accessed 24 February 2017).

Mobasher, B., Jin, X. and Zhou, Y. (2004) 'Semantically enhanced collaborative filtering on the web', in Berendt, B., Hotho, A., Mladeni, D., Someren, M., Spiliopoulou, M. and Stumme, G. (Eds.): *Web Mining: From Web to Semantic Web*, Number 3209 in *Lecture Notes in Computer Science*, pp.57–76, Springer Berlin Heidelberg.

Neff, J., Dafouz, E., Herrera, H., Martínez, F., Rica, J.P., Díez, M., Prieto, R. and Sancho, C. (2003) 'Contrasting learner corpora: the use of modal and reporting verbs in the expression of writer stance', *Language and Computers*, Vol. 48, No. 1, pp.211–230.

New York Times (2017) *How to Submit an Op-Ed Article* [online] http://www.nytimes.com/content/help/site/editorial/op-ed/op-ed.html (accessed 13 April 2016).

NewVantage (2017) *Big Data Executive Survey 2017* [online] https://leadtail.com/social-insights/executive-reading-list/ (accessed 30 May 2017).

Quemada, J. and Simon, B. (2003) 'A use-case based model for learning resources in educational mediators', *Journal of Educational Technology & Society*, Vol. 6, No. 4, pp.149–163.

Salton, G., Wong, A. and Yang, C.S. (1975) 'A vector space model for automatic indexing', *Communications of the ACM*, Vol. 18, No. 11, pp.613–620.

Sanz, R.L. (2011) 'The study of authorial voice: using a Spanish-English corpus to explore linguistic transference', *Corpora*, Vol. 6, No. 1, pp.1–24.

Shim, J.P., French, A.M., Guo, C. and Jablonski, J. (2015) 'Big data and analytics: issues, solutions, and ROI', *Communications of the Association for Information Systems*, Vol. 37, No. 1, pp.797–810.

Sivarajah, U., Kamal, M. M., Irani, Z. and Weerakkody, V. (2017) 'Critical analysis of big data challenges and analytical methods', *Journal of Business Research*, Vol. 70, pp.263–286.

Stadler, C. (2015) 'Should managers read academic articles?', *Forbes* [online] http://www.forbes.com/sites/christianstadler/2015/06/16/should-managers-read-academic-articles/ (accessed 24 October 2016).

Still, K., Huhtamäki, J., Russell, M.G. and Rubens, N. (2014) 'Insights for orchestrating innovation ecosystems: the case of EIT ICT Labs and data-driven network visualisations', *International Journal of Technology Management*, Vol. 66, Nos. 2–3.

Tutty, J. (2015) 'Big data revolutionizing business in the information age', *The Courier Mail* [online] http://www.couriermail.com.au/business/qld-business-monthly/big-data-revolutionising-business-in-the-information-age/news-story/ab88f4abb859d430381a5ea10a789980 (accessed 24 November 2016).

Vanian, J. (2016) 'Has big data gone mainstream?', *Fortune* [online] http://fortune.com/2016/01/13/big-data-mainstream (accessed 24 November 2016).

Wall Street Journal (2017) *Op-Ed Guidelines for the Wall Street Journal* [online] http://www.wsj.com/articles/SB126841622758561059 (accessed 13 April 2016).

Ward, J.S. and Barker, A. (2013) *Undefined by Data: A Survey of Big Data Definitions*, arXiv preprint arXiv:1309.5821 [online] https://arxiv.org/abs/1309.5821.

Weaver, D.A. and Bimber, B. (2008) 'Finding news stories: a comparison of searches using LexisNexis and Google News', *Journalism & Mass Communication Quarterly*, Vol. 85, No. 3, pp.515–530.

Weiss, S.M. and Indurkhya, N. (1997) *Predictive Data Mining: A Practical Guide*, 1st ed., Morgan Kaufmann, San Francisco, CA.

Wyke, T. (2015) 'Academic papers are now so unintelligible that NO ONE can read them', *Mail Online* [online] http://www.dailymail.co.uk/news/article-3223513 (Accessed 11 November 2016).

Appendix 1

Topic extraction method

The coordinate of a term across a given SVD dimension in the term-by-topic matrix serves as the weight of association between the term and the topic. Let v_{il} denotes the association weight between term i and topic l. Given topic l, the SAS' proprietary algorithm uses a *term cutoff threshold*, denoted by δ_l, to determine whether a term belongs to the topic. δ_l is obtained by summing the mean and the standard deviation of the association weights for topic l such that for topic l:

$$\delta_l = \overline{v}_l + s_{v_l}, \qquad l = 1, \ldots, k$$

where

$$\overline{v}_l = \frac{\sum_{i=1}^{n} v_{il}}{n}$$

and

$$s_{v_l} = \frac{\sum_{i=1}^{n} \left(v_{il} - \overline{v}_l\right)^2}{n-1}.$$

Term i belongs to topic l if $|v_{il}| \geq \delta_l$. A term may belong to none or multiple topics.

Similarly, the elements of the document-by-topic matrix represent the degree of association between documents and topics. Let ω_{jl} denotes the degree of association between document j and topic l. The *document cutoff threshold* for each topic determines whether the topic is present in a document. Let θ_l denotes the document cutoff threshold for topic l. θ_l is the average plus one standard deviation of association weights between topic l and all m documents:

$$\theta_l = \overline{\omega}_l + s_{\omega_l}, \qquad l = 1, \ldots, k$$

where

$$\overline{\omega}_l = \frac{\sum_{j=1}^{m} \omega_{jl}}{m}$$

and

$$s_{\omega_l} = \frac{\sum_{j=1}^{m} \left(\omega_{jl} - \overline{\omega}_l\right)^2}{m-1}.$$

Document j contains topic l if $|\omega_{jl}| \geq \theta_l$. A document may cover none or multiple topics.

Appendix 2

Topics' descriptive terms

Topic	*Descriptive terms**									
Evidence-based decision making	analytics	+insight	data	big data	+analysis	unstructured	+tool	+organisation	predictive	+decision
Big data start-ups	+start-up	+software	+firm	+company	+investor	tech	+cloud	+venture	+fund	+capital
Digital economy	+digital	+organisation	+cloud	+transformation	+economy	artificial	+skill	+innovation	+artificial intelligence	+mobile
Retail analytics	+consumer	+retailer	+brand	online	+customer	advertising	+shop	+retail	+shopper	+marketer
Big data research	+researcher	+science	+computer	+scientist	+research	scientific	+algorithm	+human	+search	+disease
Smart cities	+city	+vehicle	+car	+smart	+sensor	+energy	+device	+traffic	+monitor	+transportation
Cloud computing	+storage	+cloud	+server	+application	+enterprise	+workload	+network	+centre	+performance	computing
Big data ROI	+quarter	+revenue	+percent	+price	+growth	+stock	+report	earnings	+sale	+investor
Privacy and security	+privacy	+law	+security	+protection	+breach	+protect	legal	personal	+government	surveillance
Big data education	+student	+school	+university	+college	+graduate	+skill	+education	+science	+program	+faculty
China's investment on big data	China	Chinese	yuan	+province	+city	+percent	e-commerce	jpg	+government	+country
Big data in lending industry	+bank	+loan	+credit	+lender	+borrower	+lend	+financial	+payment	+mortgage	+card
Healthcare analytics	+patient	health	+care	medical	clinical	+hospital	+treatment	+disease	healthcare	+cancer
Big data job market	java	+employer	+job	+skill	+experience	+applicant	+candidate	technical	+status	employment
Smart farming	+farmer	+farm	+agriculture	+crop	+plant	+soil	+production	agricultural	+food	+yield
Electoral analytics	+election	+voter	+campaign	+vote	political	+party	+poll	+democrat	politics	+candidate

Note: (*) + indicates a parent term (i.e., the root word after normalisation).