Centre for Central Banking Studies

Text mining for central banks

David Bholat, Stephen Hansen, Pedro Santos and Cheryl Schonhardt-Bailey

BANK OF ENGLAND

Text mining for central banks

David Bholat, Stephen Hansen, Pedro Santos and Cheryl Schonhardt-Bailey(*)

david.bholat@bankofengland.co.uk
stephen.hansen@upf.edu
pedro.santos@bankofengland.co.uk
c.m.schonhardt-bailey@lse.ac.uk

Although often applied in other social sciences, text mining has been less frequently used in economics and in policy circles, particularly inside central banks. This Handbook is a brief introduction to the field. We discuss how text mining is useful for addressing research topics of interest to central banks, and provide a step-by-step primer on how to mine text, including an overview of unsupervised and supervised techniques.

(*) We thank Ayeh Bandeh-Ahmadi, Aude Bicquelet, David Bradnum, Peter Eckley, Jo Gill, David Gregory, Sujit Kapadia, Tom Khabaza, Christopher Lovell, Rickard Nyman, Paul Ormerod, Paul Robinson, Robert Smith, David Tuckett, Iulian Udrea and Derek Vallès for their input.

ccbsinfo@bankofengland.co.uk
Centre for Central Banking Studies, Bank of England, Threadneedle Street, London, EC2R 8AH

Series editor: Andrew Blake, email andrew.blake@bankofengland.co.uk

ISSN: 1756-7270 (Online)

Contents

Text mining for central banks

Introduction

Text mining (sometimes called natural language processing[1] or computational linguistics) is an umbrella term for a range of computational tools and statistical techniques that quantify text.[2] Text mining is similar to reading in that both activities involve extracting meaning from strings of letters. However, the computational and statistical analysis of text differs from reading in two important respects. First, computer-enabled approaches can process and summarise far more text than any person has time to read. And second, such approaches may be able to extract meaning from text that is missed by human readers, who may overlook certain patterns because they do not conform to prior beliefs and expectations.

Although widely applied in other fields such as political science and marketing, text mining has been historically less used as a technique in economics. This is particularly the case with respect to research undertaken inside central banks. There may be a couple of reasons why this has been the case. First, it may not be obvious that text can be described and analysed as quantitative data.[3] As a result, there is probably a lack of familiarity in central banks with the tools and statistical techniques that make this possible. Second, even if central bankers have heard of text mining, they already have access to other readily available quantitative data. The opportunity and other types of costs from transforming texts into quantitative data, and learning new tools and techniques to analyse these data, may be viewed as outweighing the expected benefits.

However, text mining may be worth central banks' investment because these techniques make tractable a range of data sources which matter for assessing monetary and financial stability and cannot be quantitatively analysed by other means. Key text data for central banks include news articles, financial contracts, social media, supervisory and market intelligence, and written reports of various kinds. With text mining techniques, we can analyse one document or a collection of documents (a corpus). A document could be a particular speech by a Bank of England (but here referred to as 'Bank') Monetary Policy Committee (MPC) member, a staff note, or a field report filed by an Agent.[4] The corresponding corpus would be all MPC member speeches, staff notes, and field reports, respectively.

Although the intentional use of text mining techniques by central banks is still limited, central bankers already do reap the benefits of text mining applications on a daily basis. Consider, for example, how often central bankers (or anyone) Google for information, or use spellcheck before publishing documents. Consider also the spam detection firewalls used to insulate central banks from cyber-attacks, or the query functionality in citation databases used for retrieving the existing scholarly literature on any given topic. In these and other instances, text mining techniques operate in the background to help central banks perform their jobs more efficiently.

The additional purpose of this Handbook then is to demonstrate the value central banks may gain from a more conscious application of text mining techniques, and to explain some of them using examples relevant to central banks. The Handbook proceeds in two main parts. The first part explains how text mining can be applied in central bank research and policymaking, drawing on examples from the existing literature. The second part of the Handbook then provides a step-by-step primer on how to mine text. We begin by explaining how to prepare texts for analysis. We then discuss various text mining techniques, starting with some intuitive approaches, such as Boolean and dictionary techniques, before moving on to discuss those that are more elaborate, namely Latent Semantic Analysis, Latent Dirichlet Allocation and Descending Hierarchical Classification.

Boolean and dictionary text mining, on the one hand, and Latent Semantic Analysis, Latent Dirichlet Allocation and Descending Hierarchical Classification techniques, on the other, map onto different epistemologies, that is, different approaches to knowledge-making: deduction and abduction, respectively.[5] Deduction starts from a general theory and then uses particular datasets to test the validity of the theory. By contrast, abduction attempts to infer the best explanation for a particular event based on some data, without ambition to generate an explanation generalisable to other cases.[6] Boolean and dictionary text mining are deductive approaches in that they start with a predefined list of words, motivated by a general

(1) Natural language processing is the computational processing and analysis of naturally occurring human languages, as opposed to programming languages, like Java.
(2) There are also computer assisted approaches for *qualitative* analysis of text. These are outside the scope of this Handbook. However, see the following link for an overview and comparison of some of the qualitative text mining tools: http://www.surrey.ac.uk/sociology/research/researchcentres/caqdas/support/choosing/. See also Upshall (2014).
(3) Text is sometimes called unstructured data, in contrast to structured data (numbers). However, referring to text as unstructured is somewhat misleading. Text does have structure; most obviously grammar, but also structural patterns of various kinds that text mining techniques extract.
(4) Agents are Bank staff scattered across the UK who provide intelligence on local economic conditions.
(5) Of course, deduction and abduction are ideal types. In reality, all explanatory approaches are mixed. Nevertheless, we think this classification helps situate different text mining techniques in terms of their similarities and differences.
(6) Induction is a third epistemology that, like abduction, starts from data without priors, but like deduction, then seeks to generate general theoretical claims.

theory as to why these words matter. The strengths of this approach are simplicity and scalability. Code for its implementation is typically just a few lines long, and can be applied easily to massive text files. However, the weakness of this approach is its focus only on words pre-judged by the researcher to be informative while ignoring all other words. By comparison, Latent Semantic Analysis, Latent Dirichlet Allocation and Descending Hierarchical Classification infer thematic patterns in a particular corpus without claiming that these patterns hold in other documents. The main strength of these techniques is that they analyse all words within the sampling frame and yield more sophisticated statistical outputs. Their main disadvantage is programming complexity.

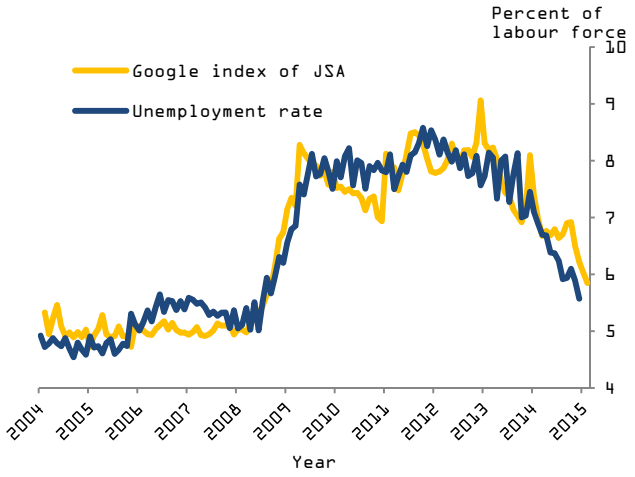
Text mining is a vast topic. By necessity we have had to be selective in the techniques we cover in the Handbook. We mostly focus on unsupervised machine learning techniques. Unsupervised machine learning involves taking *unclassified* observations and uncovering hidden patterns that structure them in some meaningful way.(1) These techniques can be contrasted to supervised machine learning. Supervised machine learning starts with a researcher *classifying* observations to 'train' an algorithm under human 'supervision' – to 'learn' the correlation between the researcher's ascribed classes and words characteristic of documents in those classes (Grimmer and Stewart (2013)). While we touch on supervised machine learning briefly in conclusion, the focus of this Handbook is on unsupervised machine learning techniques because they resonate with the Bank's evolving 'big data' ethos (Bholat (2015); Haldane (2015)). Throughout the Handbook we bold key terms where they are defined, as we have done in this introduction.

1 Text as data for central bank research

In order to motivate our discussion of text mining, we start by considering current core research topics of interest to central banks, using the Bank's recently released One Bank Research Agenda (OBRA) Discussion Paper as a proxy (Bank of England (2015)). Indeed the OBRA Discussion Paper identifies text as a data source whose analytical potential has not been fully tapped. This is particularly so given that there is a wealth of new text data available via social media and internet search records. McLaren and Shanbhogue (2011) offer a fine example of what can be done. Using Google data on search volumes, they find that such data provides a timelier tracking of key economic variables than do official statistics. For instance, Figure 1 shows that Google searches for Jobseeker's Allowance (JSA)(2) closely track official unemployment.

One issue that interests central banks is measuring risk and uncertainty in the economy and the financial system. A recent contribution in this direction is research by Nyman et al. (2015). Nyman and his co-authors start from a general theory – the emotional finance hypothesis. This is the hypothesis that individuals gain conviction to take positions in financial markets by creating narratives about the possible outcomes of their actions. These conviction narratives embody emotion such as excitement about expected gains, and anxiety about possible losses. According to Nyman and his co-authors, these narratives are not composed by individuals in isolation. Rather, they are constructed socially, through interactions like when individuals talk to one another. Through these social interactions, narratives are created and disseminated, with potential impact on asset prices.

Figure 1 Googling the labour market



Source: McLaren and Shanbhogue (2011).

They test their hypothesis by looking at three text data sources: the Bank's daily market commentary (2000-2010), broker research reports (2010-2013) and the Reuters' News Archive (1996-2014). Sentiment is measured by constructing the sentiment ratio in Equation 1.

Equation 1 Sentiment ratio(a)

$$SI[T] = \frac{(|\text{Excitement}| - |\text{Anxiety}|)}{|T|}$$

Source: Nyman et al. (2015).

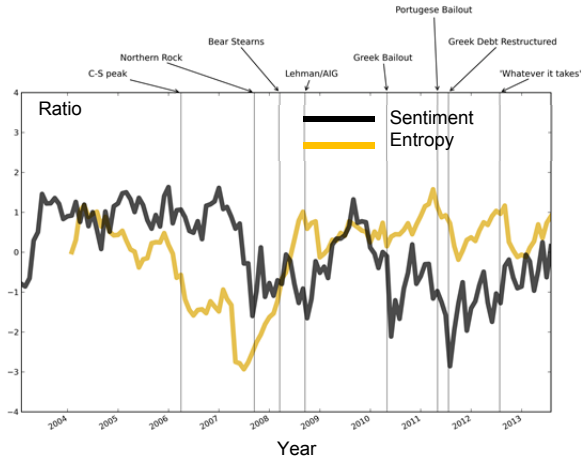
(a) *SI[T]* is the sentiment ratio of document *T*, |Excitement| is the number of 'excitement' words, |Anxiety| is the number of 'anxiety' words and |*T*| is the total number of words in document *T*.

The sign of the ratio gives an indication of market sentiment: bullish, if the ratio is positive, or bearish, if the number is negative. The ratio is then compared with historical events and other financial indicators.

In addition, they measure narrative consensus. In particular, their approach is to group articles into topic clusters.(3) The uncertainty in the distribution of topics then acts as a proxy for uncertainty. In other words, reduced entropy in the topic distribution is used as an indicator of topic concentration or consensus. Figure 2 depicts the time series for the sentiment index and the consensus measure. The authors find evidence of herding behaviour (reduced entropy) and increased

excitement ahead of the recent financial crisis.

Figure 2 Sentiment and entropy in Reuters' News Archive

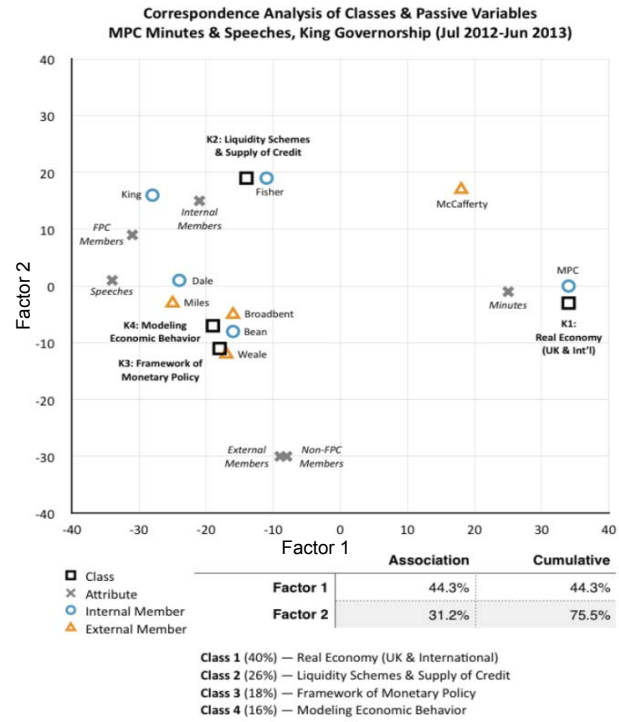


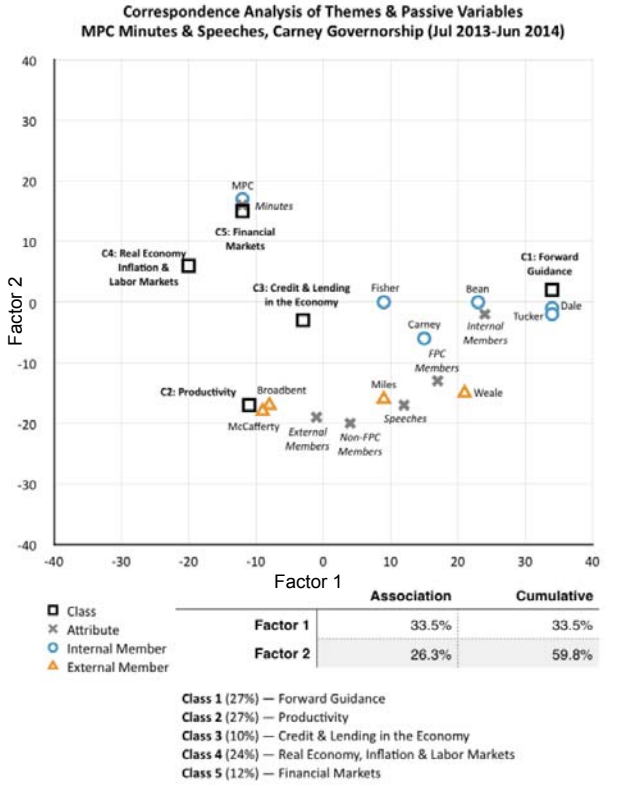
Source: Nyman et al. (2015).

Once uncertainty in the economy is measured, central banks aim to manage it. This is one of the main motivations behind the Bank's recent policy of forward guidance (Carney (2013)), by which the Bank steers expectations about the future direction of policy through communications of future intentions and official forecasts. Text mining can help here and in other similar instances to measure the extent to which Bank officials are communicating a consistent message to the outside world.(1) And assessing the efficacy of the Bank's communications is an area of research identified by the OBRA Discussion Paper.

Figure 3 from Vallès and Schonhardt-Bailey (2015) exemplifies the kind of research that can be conducted. The figure depicts the thematic content of MPC speeches and minutes in the last year of Mervin King's Governorship and the first year of Mark Carney's Governorship.

Figure 3 Thematic content of MPC Minutes(a)





Source: Vallès and Schonhardt-Bailey (2015).

(a) These graphs depict the correlations between topics and speakers in the King and Carney Governorships. The positions of the points and the distance between points reflects the degree of co-occurrences. The axes identify the maximum amount of association along factors, as explained in greater detail in Section 2.

Each graph spatially represents co-occurrences – that is, the convergence and divergence of individuals in speaking about certain topics. Spatial proximity suggests a greater degree of co-occurrence. For instance, in both graphs, the 'Real Economy' topic

(1) The outputs of algorithms for unsupervised machine learning can be used as inputs into econometric models for predicting some variable of interest, but this is a different approach from intentionally choosing the dimensions of content based on their predictive ability.
(2) Unemployment benefit paid by the government of the United Kingdom.
(3) In particular, the authors use X-means clustering algorithm, which employs Bayesian Information Criteria (BIC) to detect the optimal number of clusters. They then use Shannon entropy as a measure of the topic distribution. Increased consensus is gauged through (1) reduction in the number of topic clusters, when the actual size of each cluster remains unchanged and (2) relative growth of one particular topic, for a fixed number of topic clusters.

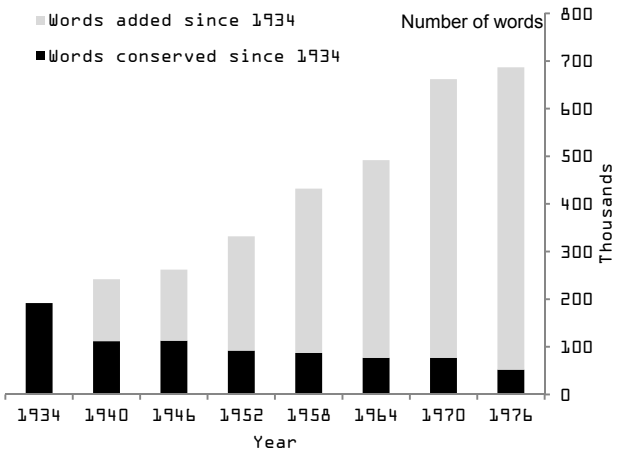
class is closely associated with the MPC when it speaks as a committee in its minutes.

A distinct divide can be detected in the topics discussed by MPC members in their external speeches during the Carney era. While some members used their speeches to discuss forward guidance, others did not. Vallès and Schonhardt-Bailey thus shed light on where the committee as a whole convey one message, while individual members are delivering more varied messages.(2)

Just as central banks want to understand whether they are communicating a consistent message, they are equally interested in whether the various policies they enact are complementary or conflicting. In fact, the OBRA Discussion Paper identifies understanding the interactions between monetary, macro-prudential and micro-prudential policy as an important research topic for the Bank. In order to understand these interactions, text mining may be useful. Here we draw inspiration from a recent paper by William Li and his co-authors titled “Law is Code” (Li et al. (2015)). Their paper tracks the increasing complexity of American law over time by analysing the complete United States (US) Code from 1926 up to the present. Striking a chord similar in tone to Haldane’s (2012) critique of complex financial regulation, the authors argue that the increasing complexity of legal code makes it difficult to understand, generates negative unintended consequences, and is a potential drag on productivity. In order to capture empirically the increasing complexity of the US Code, the authors produce several text-based metrics. These include:

1 Metrics assessing the size and substance of the Code over time. The authors interpret the lengthening of the Code as measured by word count to mean it is increasingly burdensome. They note that the gross size of the Code and its rate of growth have been increasing in recent decades. In addition, they track changes in the substance of specific sections of the Code by comparing the words added and deleted across time. For instance, Figure 4 shows changes in Title 12 (Banks and Banking) of the Code between 1934 and 1976.

Figure 4 Words conserved and added to Title 12



Source: Li et al. (2015).

2 Measures of cyclomatic complexity. In the context of their paper, cyclomatic complexity refers to the count of conditional statements in the Code, that is, clauses starting with ‘if’, ‘except’, ‘in the event’ and other such hedging words. The authors argue that conditional statements create uncertainty about when laws will be applied and therefore indicate increased complexity.

3 Measures of interconnectedness. As in the financial system, interconnectedness in the law can spawn spillovers and unintended consequences (Gai et al. (2011)). Li and his co-authors measure legal interconnectivity by looking at citations between different parts of the Code. Those sections that contain many references, and are referenced elsewhere, are ‘systemically important’ nodes in the legal network. Table 1 sets out the most important sections of Title 12. This analysis identifies the sections of Code that, if changed, would have a significant impact on other aspects of the law.

There are obvious applications of the approach taken by Li and his co-authors to issues that matter to central banks. For example, measures of textual interconnectedness could be undertaken ahead of any proposed changes to regulation or regulatory reporting forms. This could help quantify *ex ante* the potential adverse interactions between monetary, macro-prudential and micro-prudential changes. More generally, metrics tracking changes in the size, substance, cyclomatic complexity and interconnectedness of regulation could be calculated in the UK and other contexts to test the hypothesis that financial regulation has become more complex over time.

Table 1 Sections of Title 12 of the US Code with greatest interconnectivity

(1) See Rosa and Verga (2006), Blinder et al. (2008), Jansen and Haan (2010) and Bennani and Farvaque (2014) for similar investigations into the consistency of central banks’ communication. However, consistency in communication is not always good. For example, Humpherys et al. (2011) developed models to identify fraudulent financial statements from management communications and found evidence that fraudulent statements are likely to contain less lexical diversity.

(2) Other recent papers using text mining to understand central banks’ communications include analysis by Bulir et al. (2014) of central banks’ inflation reports, and analysis by Siklos (2013) of minutes from five central banks, showing how their diction changed after the financial crisis. Also Nergues et al. (2014) derive network metrics to investigate changes in discourse in the European Central Bank before and after the financial crisis.

Section Number	Name
1841	Bank Holding Company Act Definitions
101	Repealed (delivery of circulating notes)
1818	Termination of Status as Insured Depository Institution
1709	Insurance of Mortgages
1813	Federal Deposit Insurance Act Definitions

Source: Li et al. (2015).

2 Primer on text mining techniques

The first section of this Handbook was suggestive of the promise text mining holds for central banks. In this section we enumerate key steps in any text mining project, and give an overview of some specific text mining techniques. All of the techniques described in this section are a kind of content analysis summarising what texts are about. They do this by counting the number of words in a corpus. The underlying intuition is that the frequency of words and their co-occurrence are good gauges of the topic or sentiment expressed in texts.

To make this intuition more concrete, consider the words clouds in Figure 5 on the next page. The word cloud on top is derived from Governor Carney's opening remarks at the launch of the Bank's OBRA conference (Carney (2015)), while the one on the bottom is derived from the *Financial Times* (*FT*) story about the same event (Giles (2015)).

As can be seen from Figure 5, problems arise with high and low frequency words. For example, the words 'the' and 'and' occur very often, and for this reason do little to help us distinguish one document from another. Conversely, lots of words occur only once.

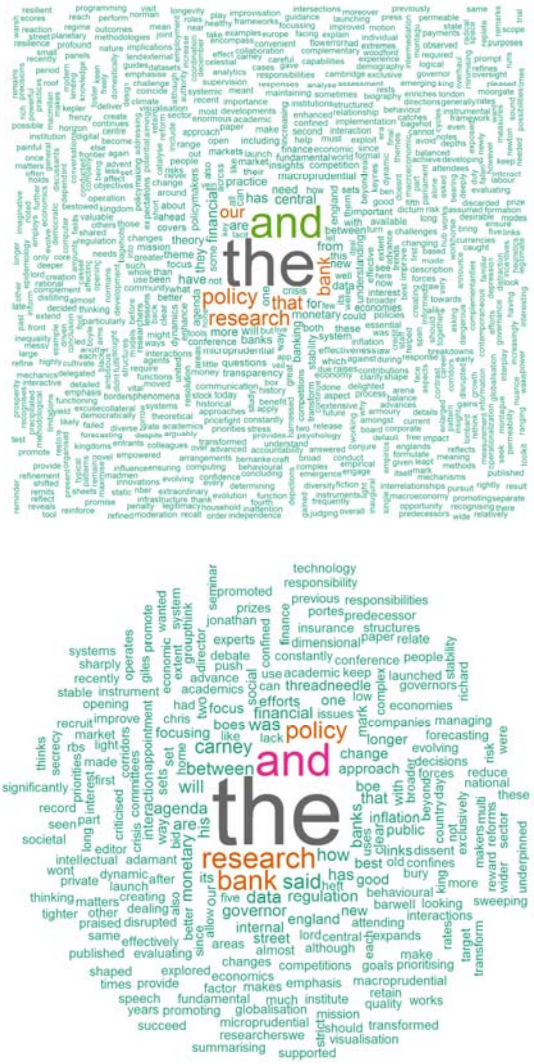


Figure 5 The Bank's OBRA conference: Mark Carney's remarks (top) and *FT* story (bottom)

Figure 6 simplifies the word clouds from Figure 5 by only showing words that occur at least twice in the texts, and by removing the words 'the' and 'and'. The two clouds are similar in content with the exception that Mark Carney did not speak of himself in the third person so "Mr Mark Carney" is absent from the cloud on top.

This example teaches two lessons about text mining to which we will return again and again in this Handbook. First, some (visually and statistically) prominent words may reveal nothing significant beyond the obvious, e.g. Carney did not use his own name in his speech. And second, we may want to omit certain common words, such as 'the' and 'and', from our sampling frame because they do not add analytical value. In other words, we can remove words in the frequency distribution without loss to understanding what a corpus is about.

Figure 6 Simplified word clouds: Mark Carney's remarks (top) and *FT* story (bottom)

Table 2 represents these word clouds as a term-document matrix (TDM), with rows corresponding to unique terms and columns corresponding to each document. Element $x_{i,j}$ represents the count of the i^{th} term in document j.

Table 2 Matrix representation of word clouds

Document Term	Mark Carney's remarks on Bank's Research Agenda	*FT* story on BoE Research Agenda
policy	37	8
research	35	11
bank	28	10
that	28	3
our	25	2
financial	18	3
central	16	1
thinks	0	1
uses	0	1
wanted	0	1
won't	0	1

Two general features of term-document matrices such as Table 2 are that they tend to be high-dimensional and sparse. Any one document will contain only a subset of all unique terms, and the rows corresponding to unused terms will all be zero. For example, the full term-document matrix for the OBRA related texts has 906 dimensions (or unique terms) and 42% of the word counts are zero. The key task then becomes how to extract low-dimensional information from documents that are high-dimensional by nature. This is analogous to a situation in which a researcher has a database with thousands of covariates and is attempting to choose which subset of them, or which summary statistics, should be included in regression analysis.

So while the basic intuition behind using word counts as an indicator of texts' content is simple, actual implementation of this approach is more complicated. In the next subsections, we focus on techniques to reduce dimensionality – the number of words or variables – and sparseness.

Analytical pre-processing

The first step in text mining is to define the corpus in scope. If the ambition is to generalise from the corpus to a larger population of documents, then standard sampling rules apply. Documents should be representative and selected using random or some other probabilistic sampling strategy. One challenge sometimes encountered at this stage is duplication – that there may be more than one instance of the same document in a corpus. For example, in large newspaper databases, the same article sometimes occurs more than once, perhaps because there are different editions of a newspaper in different countries. Such duplication can distort inference by over-representing certain documents. De-duplication can be accomplished by manually reviewing the corpus to ensure each document is a unique observation, or by using an algorithm as in Eckley (2015).

Once the corpus is specified, the next step is to transform it into a format amenable to analysis. Making it so is often cumbersome and time consuming. For instance, documents saved in pdf format may need to be converted and saved as text (txt) files. Also, if the texts exist only in paper form, like many archival artefacts, then they need to be scanned and converted into txt files using optical character recognition software.

Once texts are transformed into an appropriate format, the next step is to break documents into tokens. This step involves representing text as a list of words, numbers, punctuation, and potentially other symbols like currency markers or copyright signs.[1] Again in theory this sounds simple, but in practice it is much more difficult.

Consider an extract from one sentence of the Bank's mission statement: *Promoting the good of the people of the United Kingdom*. The number of words in this sentence could be counted in at least two ways. One way is to count each of the discrete words or tokens. Using this approach, the extract has ten tokens. But another way to count the number of words in a

sentence is to count distinct words. So while there are ten tokens in the extract, there are just seven distinct words, because the token '*the*' occurs three times and 'of' appears twice. We can represent the number of word types as a vector [1,3,1,2,1,1,1], as we have done in Table 3. In text mining, vector representations of text are called bag-of-words representations.

Table 3 A vector representation of words

Document / Term	Bank's mission statement (extract)
Promoting	1
the	3
good	1
of	2
people	1
United	1
Kingdom	1

Note that the vector representation in Table 3 treats 'United' and 'Kingdom' as separate words. However, another representation, arguably more reflective of the meaning or semantics of the sentence, might treat these two tokens as an instance of a single concept i.e. the 'United Kingdom.'

Table 4 A vector representation of tokens

Document / Term	Bank's mission statement (extract)
Promoting	1
the	3
good	1
of	2
people	1
United	1

Writing an algorithm to tokenise text so that it always conveys the correct meaning is difficult. For example, if an algorithm represented every instance where 'United' is followed by the word 'Kingdom' as an instance of the term 'United Kingdom', then the algorithm would incorrectly treat 'united' and 'kingdom' in the following sentence as one token instead of two: *The marriage of Isabella of Castile to Ferdinand of Aragon created a united kingdom in Spain.*

Besides exemplifying the challenges of tokenising text so word counts accurately convey meaning, the sentence above also shows that unadjusted word counts may be a poor indicator of a document's distinctive content.

(1) One way to tokenise text is to use a regular expression function which allows researchers to search for patterns in text and split the text on these patterns. For example, a search for the words 'Bank of England' might use a regular expression that searches for a word that begins with the letter 'B' followed by: three alphabetical characters, a space, the word 'of', a space and a word that begins with the letter 'E'.

In the above example, the word '*the*' is the mode. So, as noted at the start of this section, a key aspect of text mining is to reduce the dimensionality of bag-of-word representations to eliminate 'noise' and hone in on documents' distinctive content. There are a number of techniques available to deal with words that are superfluous to the content of the corpus. Some of them are listed below.

1 *Strip out punctuation and rare words*. 'It', 'the', 'a' and the like are extremely common words but contribute little to distinguishing the content of one document from the content of another. So these so-called stop-words are often dropped from the sampling frame.(1) For example, if we dropped the words 'of' and 'the', then the extract of the Bank of England's mission statement could be represented as follows:

Table 5 A vector representation using stop-words

Document / Term	Bank's mission statement (extract)
Promoting	1
good	1
people	1
United	1

2 *Recast words into their common linguistic root*. Another procedure for reducing dimensionality is lemmatisation. Lemmatisation uses Part-of-Speech (POS) tagging to identify the grammatical class (part-of-speech) to which each word belongs – nouns, pronouns, verbs, etc. – and convert them into their base form. For example, 'saw' tagged as a verb would become 'see' but not when tagged as a noun. POS tagging is difficult because words often belong to more than one class. For example, the word 'bank' can be used both as a noun and as a verb. In instances where the POS for a word is multiple, parsing algorithms can be used to detect the correct POS based on neighbouring words.(2) Consider the phrase: *Fed increases interest rate* (Jurafsky and Manning (2012)). Each one of these words can be either a noun or a verb. For example, the word 'Fed' may refer to the Federal Reserve or signify the past tense of 'feed.' A parsing algorithm would disambiguate each of these tokens by reference to one another, until each token was assigned to a grammatical class.

3 *Stemming.* In practice, many text miners simply stem words because this procedure tends to be faster and easier to implement than lemmatisation.(3) Stemming involves cutting off affixes and counting just stems. For example, the word 'banking' contains the stem 'bank' and the affix '-ing'. Therefore, 'banking' and

'bank' once stemmed would be treated as two instances of the same token.[4]

4 *Case folding*. Case folding involves converting all alphabetic tokens to lowercase. While in some cases this might obscure meaning for some proper nouns – for instance, an acronym such as 'US' (= United States) is erroneously converted to the pronoun 'us' – in many cases this procedure simply treats as irrelevant whether a word happens to begin a sentence. Nevertheless, there are instances where case folding can be misleading. Consider the sentence: *The Bank of England is the main regulator of XYZ bank*. Case folding would treat 'Bank' and 'bank' as instances of the same token when they are, in fact, different legal entities.

Boolean techniques

So far we have considered pre-processing steps common in any text mining research. We now turn to different types of text mining techniques.

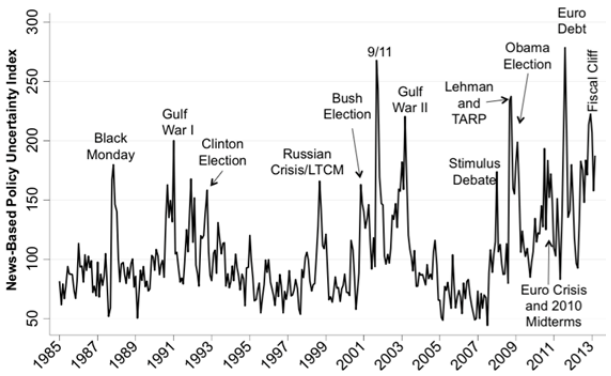
Perhaps the simplest method for mining text is to perform a Boolean search for one or more key terms. Boolean search techniques combine individual terms or phrases with logical operators like AND, OR, and NOT to form search expressions. Each document is then assigned 1 or 0 depending on whether the expression is true (= 1) or not (= 0). Since Boolean search is incorporated in most major internet search engines, one advantage of this technique is that researchers do not have to have access to the raw texts if they already have been indexed. For example, Google or Yahoo will return an estimate of the number of websites that satisfy Boolean search criteria.

An example of research relevant to central banks using Boolean techniques is the work of Baker, Bloom, and Davis (2013) who estimate economic policy uncertainty by analysing major US and European newspapers. For each paper on each day beginning in 1985, they count the number of articles containing words related to uncertainty and the economy. Specifically, the authors compute the number of articles per day that satisfy the following search criteria:

1 Article contains "uncertain" OR "uncertainty", AND
2 Article contains "economic" OR "economy", AND
3 Article contains "congress" OR "deficit" OR "federal reserve" OR "legislation" OR regulation" OR "white house".

The time series of this normalised count – scaled by the total number of articles in the newspaper each month – then acts as a proxy for policy uncertainty.[5] As can be seen in Figure 7, their index is able to capture important political and financial market events.

Figure 7 News-based policy uncertainty index



Source: Baker, Bloom, and Davis (2013).

Another example of Boolean text mining is recent work the Bank did in the run-up to the Scottish independence vote. At the time there were concerns about the negative consequences that might have ensued following a 'yes' result; for example, runs on Scottish banks (Bank of England (2014)). As poll results and public sentiment fluctuated, the Bank monitored these risks by looking at Twitter traffic. Tweets containing terms and combinations of terms suggestive of bank runs were monitored. Tweets were monitored for a full week, but with particular attention on the night of the referendum vote. In the end, there was little traffic, as the 'no' on independence vote became apparent.[1]

In conclusion, it could be argued that Boolean text-based measures of risk and uncertainty have four comparative strengths over more traditional measures like the VIX index. First, by modifying the search terms used for articles, one can capture broader uncertainty beyond those specific to financial markets. Second, the source of uncertainty is often stated explicitly in text; 'uncertainty caused by the referendum vote', for example. By contrast, with uncertainty measures based on financial asset prices, the source of uncertainty is not always clear. Third, by extending media text searches back in time, we can obtain much longer time series than those based on option prices. For example, the VIX is available only from the late 1980's, while Baker, Bloom, and Davis extend their index back to 1900. Fourth and finally, we can obtain cross-country measures of uncertainty by focusing on separate countries' media text, whereas VIX is merely a proxy of uncertainty in US equity prices, which may be a poor gauge of uncertainty in other countries.

(1) Lists of stop-words are available online, e.g. http://snowball.tartarus.org/algorithms/english/stop.txt
(2) MaltParser (Nivre et al. (2006)) is one of the most widely used.
(3) The Porter stemmer is popular for texts in English.
(4) The output after stemming may not be a naturally occurring word, without loss of interpretative meaningfulness. For example, the stem of 'inflation' is 'inflat'. As noted before, lemmatisation gives a more refined result i.e. typically returns the dictionary form of a word.
(5) See http://www.policyuncertainty.com/ for more details.

Dictionary techniques

Another set of strategies for mining text are dictionary techniques, such as those used by Nyman and co-authors, as discussed in Section 1. Dictionary techniques typically proceed in two steps:

1 Start by defining a list of key words to capture content of interest;
2 Then represent each document in terms of the (normalised) frequency of words in the dictionary.

Unlike Boolean techniques, dictionary techniques measure the intensity of word use, which may be a better measure of the corpus' content. But unlike Boolean techniques, the application of dictionary techniques requires the researcher to have access to the raw texts.

For example, let the dictionary $D = \{\text{labour, wage, employ}\}$ be the stems of various words relating to labour markets. We can then represent each document d as the share s_d of words that are in the dictionary – Equation 2.

Equation 2 Normalised frequency of words in a document

$$s_d = \frac{\left(\begin{array}{c}\text{number of 'labour' occurrences} \\ + \\ \text{number of 'wage' occurrences} \\ + \\ \text{number of 'employ' occurrences}\end{array}\right)}{\text{total words in document } d}$$

The existing literature provides several examples where dictionary techniques have been used to analyse financial and economic texts. For example, Tetlock (2007) authored a highly-cited paper that uses dictionaries (aka lexicons) to gauge the tone of the *Wall Street Journal* column "Abreast of the Market."(2) Specifically, his paper employs the Harvard IV-4 dictionaries containing word lists reflecting many classes, including positive and negative sentiment, pain and pleasure, and rituals and natural processes, among others.(3) Tetlock then counts the number of words in each day's column from 1984-1999. His first finding is that most of the variation in word counts across columns reflects swings from optimism to pessimism. His second finding is that the impact of an increase in negative news moves the next day's returns in a statistically significant way. After Tetlock (2007) many papers have used the same basic strategy of counting words of interest in financial market contexts and correlating them to asset prices (e.g. Aase (2011)).

However, as we have noted a few times already, there are many instances where raw word counts can be misleading. Consider the genre of text mining referred to as sentiment analysis, often also called opinion mining or subjectivity analysis. The goal of sentiment analysis is to detect the feeling expressed about some object or, in the jargon, target. The target of the sentiment might be a person, event, institution, or inanimate object, just to name a few possibilities. The simplest type of sentiment analysis is polarity detection, that is, a binary classification of positive and negative sentiments.

However, accurately gauging sentiment is complicated by a number of common linguistic subtleties such as negation, irony, ambiguity, idioms, and neologisms (Jurafsky and Manning (2012)). For example, consider the following paragraph:

> *One would have expected building society X to have been an excellent institution. It had a top-notch CEO and world-renowned analysts. Its services were highly rated by retail clients and its operations were efficient. Yet it was a total failure.*

Just counting the balance of positive and negative words in the above paragraph would not convey the sentiment of the passage.

Weighting words

A simple counting approach may be inappropriate because it can overstate the importance of a small number of very frequent words. This is potentially troublesome for two reasons:

1 Zipf's law, shorthand for the empirical observation that the frequency of a word is inversely proportional to its relative rank in a corpus. That means that a small difference in the relative rank of a word can translate into a big difference in terms of the actual word count. So relying on raw word counts may overstate their comparative importance.

2 If a word is used in many documents in a corpus, then its power to discriminate one document from another is less than if it only appears in a handful of documents. Yet it may be desirable to give more weight to words that appear in few documents since these words may indicate real differences in content.

To address these concerns, a common weighting scheme used in the text mining literature is term frequency−inverse document frequency (tf.idf), given by Equation 3.

Equation 3 Term frequency-inverse document frequency[a]

(1) In general, social media has properties that make it attractive as a data source (O'Connor et al. (2010); Bollen et al. (2011)). In particular, social media data is generated throughout the night, meaning that the authorities can monitor risks after markets close and before they reopen.
(2) See http://www.wsj.com/news/types/abreast-of-the-market
(3) See http://www.wjh.harvard.edu/~inquirer

$$\text{tf.idf}_{t,d} = (1 + \log f_{t,d}) \cdot \log\left(\frac{D}{df_t}\right)$$

(a) D is the total number of documents in the corpus, df_t is the number of documents in which term t appears and $f_{t,d}$ is frequency of term t in document d.

The first factor in Equation 3 is the frequency of term t in document d, giving lesser weight to words that appear more frequently. The second factor is the inverse document frequency of term t, giving greater weight to words that appear less frequently.

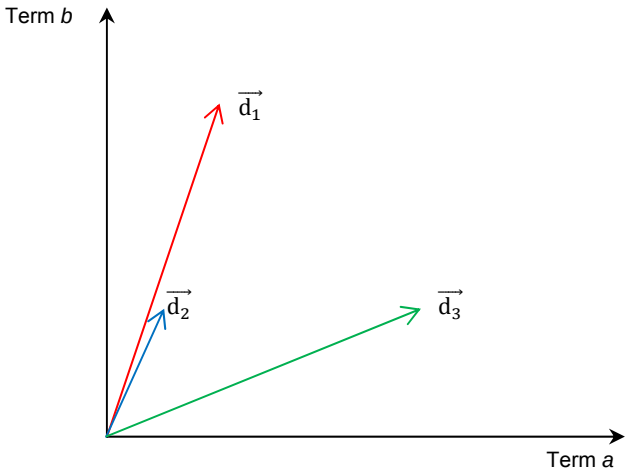
An example of tf.idf weighting is Loughran and Mcdonald (2011). Their point of departure is a critique of the Harvard IV-4 dictionaries used by Tetlock (2007). Tetlock's dictionaries contain words like 'tax', 'cost', and 'liability' that, though they convey negative sentiment in a general context, are more neutral in tone in the context of financial markets where they describe day-to-day accounting practices. Loughran and Mcdonald therefore propose a finance-specific dictionary and show they better predict asset returns than the generic dictionaries.(1) However, after tf.idf weighting, the generic dictionaries' performance improves substantially.(2) The reason is that words like 'tax' appear in many documents and so get a lower weight than other 'truly' negative words.

Vector space models

So far, we have considered techniques to identify the key topics within texts using a pre-defined set of keywords. Here we consider techniques for measuring the similarity of topics between texts.

One way to measure document similarity is to use simple Euclidean distance. For example, Kloptchenko et al. (2004) use Euclidean distance to find clusters of financial reports. However, as Figure 8 shows, this distance measure has limitations. The figure represents three hypothetical documents, each containing two terms *a* and *b*. Suppose documents 1 and 2 use terms *a* and *b* in nearly the same proportions. However, because document 1 may be much longer than document 2, their distance is quite significant. In fact, document 3, which uses term *a* relative to term *b* substantially more than document 2, would be measured as more similar to document 1, simply based on its similar length.

Figure 8 This figure represents three documents, each containing two terms *a* and *b*. Documents 1 and 2 have very similar content yet lie far apart due to differences in length



This example shows the distortions that can arise from using Euclidean distance to measure document distance. A measure that avoids these problems is cosine similarity (CS), which captures the angle formed by two vectors. Going back to Figure 8, we can see that the angle formed between documents 1 and 2 is very small – they point in the same direction since they use the two terms in nearly identical proportions. However, because the term frequencies differ for documents 1 and 3, the angle is larger. If one document vector contained only term *a* and another only term *b*, the vectors would be orthogonal. So, measuring the cosine of the angle θ formed by two documents in the vector space provides a similarity measure independent of document length. The formula for computing this is given in Equation 4.

Equation 4 Cosine similarity(a)

$$\cos\theta = \frac{\vec{d_1} \cdot \vec{d_2}}{\|\vec{d_1}\|\|\vec{d_2}\|}$$

(a) $\cdot$ is the dot product operator and $\|\vec{d_i}\|$ is the length of the vector representing document i.

An example of vector space modelling in economics is Hoberg and Phillips (2010). They take company product statements from 10-K filings with the US Securities and Exchange Commission and compute their cosine similarity. They then use this similarity score as a proxy for the companies' sector. They argue that their analysis provides a much richer and continuous measure of product substitutability than traditional industry classification codes.(1)

However, cosine similarity is no panacea. Table 6 depicts two hypothetical documents. They both are about education but use different words to communicate this. This feature of natural language is known as synonymy, meaning that the same underlying topic can be described by many different words. Even though the topic is the same in both documents, the fact that they use different vocabulary will mean their cosine similarity will be low.

(1) Available at http://www3.nd.edu/~mcdonald/Word_Lists.html
(2) Although they still find their finance-specific dictionary has greater explanatory power.

Table 6 This table shows two documents with similar content but low cosine similarity due to synonymy

(a) Document 1

school	university	college	teacher	professor
0	5	5	1	2

(b) Document 2

school	university	college	teacher	professor
10	0	0	4	0

Related to the issue of synonymy is polysemy: the fact that the same word can have multiple meanings in different contexts. Consider the two documents in Table 7.

Table 7 This table shows two documents with dissimilar content but relatively high cosine similarity due to polysemy

(a) Document 1

tank	marine	frog	animal	navy	war
5	5	3	2	0	0

(b) Document 2

tank	marine	frog	animal	navy	war
5	5	0	0	4	3

Suppose document 1 is about animals while document 2 is about war. The issue is that the same word has different meanings depending on context. For example, in a discussion of animals, 'tank' refers to a place where fish or amphibians live, while in a discussion of war it refers to a mechanised weapon. This means that essentially unrelated documents can display high cosine similarity.(2)

Latent Semantic Analysis

One assumption of vector space models is that words in text are conceptually independent of each other. However, that may not be so. For example, 'navy' and 'marine' might be conceived as surface expressions of a more fundamental, latent topic such as 'war.'

The latent variable models we now discuss take this approach. They assume words are not independent but linked together by underlying, unobserved topics. They have four virtues. They deal with synonymy by associating each word in the vocabulary to any given latent topic. They capture polysemy by allowing each word to have associations with multiple topics. They also associate each document with topics rather than words. And they allow algorithms to find the best association between words and latent variables, without using pre-defined word lists or classes, in contrast to Boolean and dictionary-based techniques.

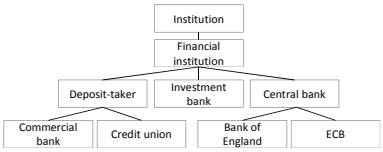
Latent Semantic Analysis (LSA) (Deerwester et al. (1990)) is one of the earliest examples of a latent variable approach. LSA begins by representing the term-document matrix by singular value decomposition (SVD). One can think of SVD as finding the principal components of the rows and the columns of a term-document matrix. That is, LSA calculates the linear combinations of terms that explain most of the variance of terms across documents, as well as the linear combinations of documents that explain most of the variance of documents across terms. The idea is then to approximate the term-document matrix using just the principal components, and to measure document similarity with the approximation rather than the true term-document matrix.(1) The hypothesis is that the principal components represent shared topics, and the discarded components represent idiosyncratic words choices.

One recent central bank relevant application of LSA is a paper by Acosta (2014), who studies the effect of greater transparency on US Federal Reserve Open Market Committee (FOMC) meetings.(2) The Fed has published verbatim accounts of FOMC meetings since October 1993. Before this date, FOMC members were not aware their deliberations were recorded. But following pressure from Congress to increase the Fed's transparency, former Fed Chairman Alan Greenspan discovered that staff had been transcribing meetings verbatim since the mid-1970s. He then agreed to release the back transcripts, and to disclose transcripts going forward with a five-year lag.

Acosta studies if greater awareness by FOMC members that their comments in meetings were being recorded and would be publicly disclosed changed their behaviour. The author applies SVD and uses the top 200 components to measure document similarity. He finds increased conformity after the publication of transcripts.

(1) See for example, the UK Standard Industrial Classification of Economic Activities 2007 (Office for National Statistics (2007)).

(2) Besides synonymy and polysemy, we can also consider the hierarchical relationship between words using thesaurus-based algorithms (Jurafsky and Manning (2012)). We say that one word is hyponym of another if the first word is a subclass of the other. For example, in the figure below, we observe that 'deposit-taker' is a hyponym of 'financial institution.' Conversely, 'financial institution' is a hypernym of 'deposit-taker.' One of the simplest approaches to assess if two words are similar is to check if they are near each other in the hierarchy. We define path length $\text{pathlen}(w_1, w_2)$ as one plus the number of edges in the shortest path in the hypernym graph between w_1 and w_2. Then, path-length similarity can be defined by $\text{pathsim}(w_1, w_2) = \frac{1}{\text{pathlen}(w_1, w_2)}$



In the figure above, the path lengths between 'commercial bank' and 'institution', and 'commercial bank' and 'investment bank' are the same. However, semantically, 'commercial bank' is more specifically related to 'investment bank' than it is to the generic idea of an 'institution.' This suggests that we need a more optimal measure. One possible improvement is to augment words in the thesaurus with Information Content (IC) values. IC values are computed based on frequency counts of words found in text. For each word *w* in the thesaurus, IC is defined as the negative log of the probability of that word. Other, more sophisticated techniques such as the Resnik (res) measure take as input two words w_1 and w_2 and output a similarity measure. The technique uses the idea of a least common subsumer (LCS), that is, the most specific word that is a shared ancestor of the two words. For example, the LCS of commercial bank and credit union is deposit-taker. The LCS is evaluated directly from the IC: $res(w_1, w_2) = IC(LCS(w_1, w_2))$.

Latent Dirichlet Allocation

One weakness of LSA is that the topics it produces are not probabilistic. Latent Dirichlet Allocation (LDA) rectifies this.

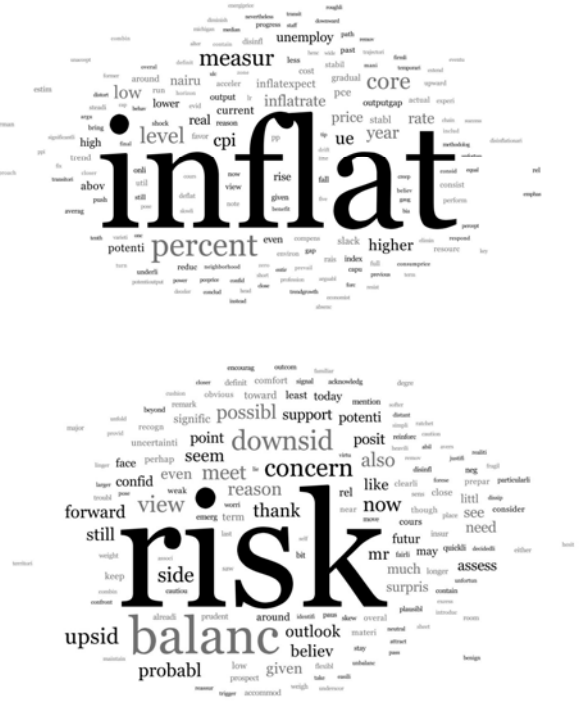
LDA is a mixed-membership model in which words and documents are assigned probabilities and related to multiple topics. This contrasts with deterministic single-membership models in which words and documents are assigned only to one topic. This probabilistic aspect of LDA is important. Suppose one imagines a topic about inflation and another about unemployment. Now consider the word 'rate.' *Prima facie* it is unclear into which topic 'rate' should go, since a topic about inflation or unemployment might feature 'inflation rate' or labour 'participation rate', respectively. Allowing probabilistic assignment of words to topics conveys this semantic flexibility.

Formally speaking, each document has its own probability distribution over topics. Then, for each word in each document, a topic assignment is made and then, conditional on the assignment, a word from the corresponding topic.(3) An example can illustrate. Suppose that speaker A and speaker B both talk about topics 1 and 2. Speaker A spends two-thirds of his time on topic 1 and speaker B spends two-thirds of her time on topic 2. In this case, A's observed words can be conceived as being chosen as follows. For each word, A draws a topic, and, with probability 0.67, this is topic 1, and with probability 0.33, this is topic 2. Once a topic is drawn for a word, the word itself is drawn from the probability distribution associated with each topic. For speaker B, the probability that a word is about topic 1 is 0.33, and about topic 2 is 0.67. But once a word is assigned to a given topic, it is drawn from that topic's distribution common across both speakers.

LDA has been widely applied in computer science and statistics, and is beginning to appear in economics. For example, Hansen and his co-authors (2014) study the effects of increased central bank transparency. Like Acosta, they find evidence of increased conformity after the release of FOMC transcripts.

Two topics they estimate are in Figure 9. The figure represents topics as word clouds in which the size of the word is proportional to its probability in the topic. No classes are imposed during estimation. These are simply the 25th and 40th word clusters estimated. Topic 25 (top) is clearly about inflation and topic 40 (bottom) is clearly about risk. So we can use the output of LDA to measure underlying content in a directly meaningful way.

Figure 9 This figure displays two topics estimated from the corpus of FOMC transcripts during the Alan Greenspan era



Source: Hansen, McMahon and Prat (2014).

Descending Hierarchical Classification

Beyond LSA and LDA, there are other text mining techniques that proceed on the basis that words are not independent of each other but reflect underlying topics/themes. For example, Schonhardt-Bailey (2013) and Vallès and Schonhardt-Bailey (2015) study central banks using the Alceste software.(1) In brief, Alceste tries to discover stable classes of terms that are maximally associated within while being minimally associated with other classes.(2) To do this, Alceste constructs a matrix that crosses all statements – so-called Elementary Context Units (ECUs)(3) – and words, where the cells indicate the presence or absence of that word in the statement, much like a term-document matrix, but at a smaller unit of textual analysis. The cells indicate the absence or presence of that word in the statement – represented by a '0' or a '1', respectively. Table 8 shows an example of such a matrix.

Table 8 Word by ECU matrix

ECU / word	1	2	3	4	5	⋯	j
1	0	1	1	1	1	⋯	0
2	1	1	0	0	0	⋯	1

(1) However there are alternative ways of performing matrix decomposition. For example, Hendry (2012) applies Non-Negative Matrix Factorization to study whether and how central banks' communication affects markets.

(2) Hendry and Madeley (2010) and Masawi et al. (2014) apply LSA to central bank documents.

(3) A more formal description of the model is as follows. Assume the corpus is made up of K topics, where an individual topic β_K is a probability vector over the V unique elements in the vocabulary. All words are therefore potentially present in all topics, albeit with different weights. At the same time, each individual document in the corpus is a probability distribution θ_d over the K topics. Each individual word $w_{d,n}$ in document d is generated via a two-step process. First, a topic allocation $z_{d,n}$ is drawn from θ_d. Second, a word is drawn from $\beta_{z_{d,n}}$. This contrasts with a single-membership model in which a document would be assigned a single topic z_d, and each word in document d would be drawn from β_{z_d}.

...	...	...	...	...	...	...	...
i	0	1	1	0	1	...	1
Totals	2	31	5	10	67	...	2

Alceste then partitions the contents of this table into two classes, with the goal of maximising the similarity of ECUs in the same class while at the same time maximising the difference between classes. The total set of ECUs in the initial matrix constitutes the first class. The algorithm then searches for a partition that minimises the number of overlapping words. Overlap is measured by the chi-square (χ^2) value of a table with two rows, comparing observed and expected distributions. The algorithm then attempts to maximise χ^2 values by repeating the partitioning process (descending hierarchical classification), that is, by testing if splitting classes into more granular subclasses improves the χ^2 values. The iterative, descending hierarchical classification process comes to conclusion when a predetermined number of iterations no longer results in statistically significant divisions.

Moreover, for each class, a list of words is produced and the strength of association between each word and the class is expressed by a χ^2 value and phi (ϕ) coefficient, where the observed distribution of words is compared with an expected one. If, for instance, the vocabulary is different in the two classes, the observed distribution will deviate systematically from an expected distribution consisting of word independence. Relationships between and among classes can also be decomposed and examined spatially using factor Correspondence Analysis. An example of correspondence analysis output was earlier shown in Figure 3. The positions of the points is contingent on correlations, where distance reflects the degree of co-occurrence.[4] With respect to the axes, correspondence analysis aims to identify a maximum amount of association along the first (horizontal) axis. The second (vertical) axis seeks to account for a maximum of the remaining association.[5]

Supervised machine learning

Latent Semantic Analysis, Latent Dirichlet Allocation and Descending Hierarchical Classification are examples of unsupervised machine learning algorithms. These can be contrasted with supervised machine learning algorithms that start with a researcher manually classifying training data with predefined classes, as in dictionary-based methods. In order to avoid the issue of over-fitting, the algorithm is then validated on another set of documents termed test data before being applied to the rest of the corpus (Grimmer and Stewart (2013)).

Perhaps the most fruitful application of supervised learning techniques in economics is when the researcher has well-motivated text classes.[6] One potential application in central banking is to associate text with a hawkish or dovish

stance, building on Apel and Grimaldi (2012).[1] A well-known application in economics using text data is Gentzkow and Shapiro (2010). Their training data is a large sample of US Congressional speeches. Each speech carries a class corresponding to the party of the speaker and identifies partisan phrases. They then score another corpus composed of press articles as left- or right-wing, based on the presence or absence of partisan phrases.

A popular supervised machine learning algorithm is Naïve Bayes.[2] Naïve Bayes makes use of Bayes' rule that the most likely class for a document is the class that maximises the product of two factors $P(c)$ and $P(d|c)$ where d is the document and c the class. The factor $P(c)$ is called the prior probability of the class. It captures how often a class occurs in the training data.

(1) Alceste is an acronym for *Analyse des Lexèmes Co-occurents dans les Énnoncés Simples d'un Texte* (Analysis of the Co-occurring Lexemes within Simple Statements of Text). The software is distributed by Image-Zafar. See: http://www.image-zafar.com/. The software was originally developed by Max Reinert. His various publications over twenty years, largely in French, document the early development of Alceste ((Reinert (1983); Reinert (1987); Reinert (1990); Reinert (1993); Reinert (1998); Reinert (2003)). An open-source, R-based reproduction of Alceste is available in the Iramuteq software (http://www.iramuteq.org/). Since 1983, an increasingly interdisciplinary community of researchers and text analysts have adopted Alceste as a text mining technique (Noel-Jorand et al. (1995); Lahlou (1996); Jenny (1997); Noel-Jorand et al. (1997); Brugidou (1998); Guerin-Pace (1998); Bauer (2000); Brugidou (2000); Brugidou (2003); Noel-Jorand et al. (2004); Schonhardt-Bailey (2005); Schonhardt-Bailey (2006); Bara et al. (2007); Schonhardt-Bailey (2008); Schonhardt-Bailey et al. (2012); Weale et al. (2012); Schonhardt-Bailey (2013); Vallès and Schonhardt-Bailey (2015)). Peart (2013) uses Alceste to study whether MPC members' preferences are stable over time.

(2) A simpler approach would be to simply count the number of occurrences and co-occurrences. For example, Ronnqvist and Sarlin (2012) investigate the co-occurrence of Finnish bank names in a major online financial forum. They turn bank names and their co-occurrences in the same post into a network, where a node's size and edge's weight are given by the number of occurrences and co-occurrences, respectively. Although it is possible to perceive some changes in concentration and strength of the connections, a more objective approach is to capture those temporal changes using metrics of network centrality. The authors detect a surge in the number of times Finnish banks are mentioned together during and after the financial crisis.

(3) Elementary Context Units ("ECUs") or statements are 'gauged sentences', which the program automatically constructs based upon the punctuation in the text. In text analysis, a persistent and difficult issue concerns the optimal length of a 'statement.' Language could be analysed in terms of sentences, paragraphs, pieces of sentences, and so on. Alceste resolves this issue by not trying to identify directly the statement length. Rather, it produces classifications that are independent of the length of the statements. Two classifications are created, each using units of different lengths of context units; only the classes which appear in both classifications are retained for analysis, and these classes are independent of the length of statements. This leaves a number of ECUs unclassified; thereby approximating a measure of goodness-of-fit. The quality of the partitioning is measured by constructing a table that crosses all the classes obtained in the first classification and all the classes obtained in the second classification. The result is a "signed χ^2 table" – that is, a data table with the positive and negative links between the classes. This table helps to select the classes which share the higher number of ECUs.

(4) To do this, correspondence analysis uses the "chi-squared distance", which resembles the Euclidean distance between points in physical space. Each squared difference between coordinates is divided by the corresponding element of the average profile (where the profile is a set of frequencies divided by their total). The reason for using the chi-squared concept is that it allows one to transform the frequencies by dividing the square roots of the expected frequencies, thereby equalizing the variances. This can be compared to factor analysis, where data on different scales are standardised.

(5) Many cases, however, require more than two dimensions to capture the dimensionality of the data. Alceste thus reports the percentage accruing to each dimension, but limits the graphical representation to two and three dimensions.

(6) Alternatively, classes may be already assigned as part of the metadata.

The factor $P(d|c)$ is called the likelihood. It captures the probability of a document d given the class, when d can be represented as a vector of words $d = x_1 + x_2 + x_3 + \cdots + x_n$, where n is the total number of words. For each word the likelihood factor can be estimated by looking at the number of times that word appears in that particular class as a ratio of all the words associated with that class in the training data. In practice these probabilities are calculated by bringing all texts within a particular class into a single combined training document for the class and then counting relative frequencies of w_i as a ratio of the overall number of words w in the training data.(3)

An example of research using Naïve Bayes is a paper by Moniz and Jong (2011) studying the effects of the Bank's MPC Minutes on future interest rate expectations. The authors employ Naïve Bayes in conjunction with other text mining techniques discussed in this Handbook. First, they look at the words on Wikipedia pages on 'Central Banking' and 'Inflation' and capture words associated(4) with 'economic growth', 'prices', 'interest rates' and 'bank lending'. These words are then used as classes in a Naïve Bayes model assigning them to sentences in the MPC Minutes. The assigned classes are then used to construct a sentiment index using a simple dictionary-based approach.(5) Finally, the LDA algorithm is employed to detect words that may act as intensifiers and diminishers of sentiment, for example 'increase' and 'moderated', respectively. Moniz and Jong's multi-method approach highlights an important point about text mining in practice – supervised and unsupervised approaches are often complementary and employed at different stages of the text mining process.

Conclusion

The purpose of this Handbook has been to demonstrate the additional value central banks can gain from applying the various text mining techniques available, and to illustrate how these techniques can be used to inform policymaking and address key research topics of interest to central banks. In closing, we wish to note that the promise of text mining for central banks is not just hypothetical but demonstrated. For example, Kevin Warsh (Warsh (2014)) cited the text mining literature (Schonhardt-Bailey (2013); Hansen, McMahon and Prat (2014)) as important influences on his review's ultimate policy recommendations for the Bank to disclose more information on its deliberations. Warsh remarked that while "studies seeking to make sense of millions of spoken words" are "daunting and imperfect", text mining has "meaningfully advanced our understanding" of central banks (Warsh (2014)). More generally, this Handbook has shown how text mining can be a useful addition to central banks' analytical arsenal and help them achieve their policy objectives.

(1) However, it may be the case that it is difficult to classify documents beforehand (Grimmer and Stewart (2013)). For example, the multi-faceted nature of meetings, speeches and conversations may make it difficult to reduce a document to a single topic.

(2) Naïve Bayes is 'naïve' in two senses. First, it starts from the simple bag-of-words assumption that word order does not matter so it only considers the frequency of words in a document. Second, it assumes that the probability of each word appearing in a given class is independent of the presence of other words, even though we have already noted this is likely to a mistaken assumption (Jurafsky and Manning (2012)).

(3) Of course, it is possible that a word associated with a class does not appear in the training data. Suppose bank X is headquartered in London. However, suppose the word 'London' does not appear in a 2,000 word training document classified as 'bank X'. As a result, the estimated likelihood of London given the topic of bank X would be zero. In order to avoid this spurious result, we can use a procedure called Laplace add-1 smoothing, which simply involves adding 1 to the following equation:

$$\hat{P}(w_i|c_j) = \frac{\text{count}(w_i, c_j)+1}{\sum_{w \in V} \text{count}(w, c_j)+V}$$

where w_i is a particular word, c_j is a particular class and V represents all words in the corpus.

(4) The authors use *TextRank*, a graph-based ranking algorithm to detect clusters of associated words (Mihalcea and Tarau (2004)).

(5) The authors use the General Inquirer dictionary http://www.wjh.harvard.edu/~inquirer/, as discussed previously.

References

Aase, K G (2011), 'Text Mining of News Articles for Stock Price Predictions', MSc Thesis, Norwegian University of Science and Technology.

Acosta, J M (2014), 'FOMC Responses to Calls for Transparency: Evidence from the Minutes and Transcripts Using Latent Semantic Analysis', Honours Thesis, Department of Economics, Stanford University. Available at http://economics.stanford.edu/content/honors-thesis-2014

Apel, M and Grimaldi, M (2012), 'The Information Content of Central Bank Minutes', *Sveriges Riksbank Working Paper Series*, No. 261.

Baker, S R, Bloom, N and Davis, S J (2013), 'Measuring Economic Policy Uncertainty', *Chicago Booth Research Paper 13-02*.

Bank of England (2014), 'Inflation Report Q&A 13th August 2014'. Available at http://www.bankofengland.co.uk/publications/Documents/inflationreport/2014/conf130814.pdf

Bank of England (2015), 'One Bank Research Agenda Discussion Paper'. Available at http://www.bankofengland.co.uk/research/Documents/onebank/discussion.pdf

Bara, J, Weale, A and Biquelet, A (2007), 'Analysing Parliamentary Debate with Computer Assistance', *Swiss Political Science Review*, Vol. 13, No. 4, pages 577—605.

Bauer, M (2000), 'Qualitative Researching with Text, Image and Sound: A Practical Handbook', In *Classical Content Analysis: A Review*, by M.W. Bauer and G. Gaskell, *Sage Publications*, London, pages 131—151.

Bennani, H and Farvaque, E (2014), 'Speaking in Tongues? Diagnosing the consistency of central banks' official communication'. Available at http://www.econ.cam.ac.uk/epcs2014/openconf/modules/request.php?module=oc_program&action=view.php&id=198

Bholat, D (2015), 'Big data and central banks', *Bank of England Quarterly Bulletin,* Vol. 55, No. 1, pages 86—93.

Blinder, A S, Ehrmann, M, Fratzscher, M, De Haan, J and Jansen, D J (2008), 'Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence', *ECB Working Paper Series,* No. 898.

Bollen, J, Mao, H and Zeng, X (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science*, Vol. 2, No. 1, pages 1—8.

Brugidou, M (1998), 'Epitaphes, l'image de Francois Mitterrand à travers l'analyse d'une question ouverte posée à sa mort (Epitaphs, Francois Mitterrand's Image: An Analysis of an Open Question Asked on His Death)', *Revue Française de Science Politique*, Vol. 48, No. 1, pages 97—120.

Brugidou, M (2000), 'Les discours de la revendication et de l'action dans les éditoriaux de la presse syndicale (1996-1998) (The Discourse of Demands and Action in Trade Union Press Editorials (1996-1998))', *Revue Française de Science Politique*, Vol. 50, No. 6, pages 962—992.

Brugidou, M (2003), 'Argumentation and Values: An Analysis of Ordinary Political Competence Via An Open-Ended Question', *International Journal of Public Opinion Research*, Vol. 15, No. 4, pages 413—430.

Bulir, A, Cihak, M and Jansen, D-J (2014), 'Does the Clarity of Inflation Reports Affect Volatility in Financial Markets?', *IMF Working Paper*, No. *14/175.*

Carney, M (2013), 'Crossing the threshold to recovery', Bank of England Speech, 28 August.

Carney, M (2015), 'One Bank Research Agenda: Launch Conference', Bank of England Speech, 25 February.

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, pages 391—407.

Eckley, P (2015), 'Measuring economic uncertainty using news-media textual data', *MPRA Paper No. 64874.* Available at http://mpra.ub.uni-muenchen.de/64874/

Gai, P, Haldane, A and Kapadia, S (2011), 'Complexity, concentration and contagion', *Journal of Monetary Economics,* Vol. 58, No. 5, pages 453—470.

Gentzkow, M and Shapiro, J M (2010), 'What drives media slant? Evidence from U.S. daily newspapers', *Econometrica,* Vol. 78, No. 1, pages 35—71.

Giles, C (2015), 'Bank of England Mark Carney expands Research Agenda', *Financial Times 25th February 2015*.

Grimmer, J and Stewart, B M (2013), 'Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts', *Political Analysis*, Vol. 21, pages 267—297.

Guerin-Pace, F (1998), 'Textual Analysis, An Exploratory Tool for the Social Sciences', *Population: An English Selection, special issue of New Methodological Approaches in the Social Sciences*, Vol. 10, No. 1, pages 73—95.

Haldane, A (2012), 'The dog and the Frisbee', Bank of England speech, 31 August.

Haldane, A (2015), 'The promise of new data and advanced analytics', Bank of England speech, 25 February.

Hansen, S, McMahon, M and Prat, A (2014), 'Transparency and Deliberation within the FOMC: a Computational Linguistics Approach', *CEP Discussion Papers* DP1276, Centre for Economic Performance, LSE.

Hendry, S (2012), 'Central Bank Communication or the Media's Interpretation: What Moves Markets?', *Bank of Canada Working Paper 2012-9*.

Hendry, S and Madeley, A (2010), 'Text Mining and the Information Content of Bank of Canada Communications', *Bank of Canada Working Paper 2010-31*.

Hoberg, G and Phillips, G M (2010), 'Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis', *The Review of Financial Studies*, Vol. 23, No. 10, pages 3773—3811.

Humpherys, S, Moffitt, K C, Burns, M B, Burgoon, J K and Felix, W F (2011), 'Identification of fraudulent financial statements using linguistic credibility analysis', *Decision Support Systems*, Vol. 50, pages 585—594.

Jansen, D J and De Haan, J (2010), 'An Assessment of the Consistency of ECB Communication using Wordscores', De *Nederlandsche Bank Working Paper 259.*

Jenny, J (1997), 'Techniques and formalized practices for content and discourse analysis in contemporary French sociological research', *Bulletin de Méthodologie Sociologique*, Vol. 54, pages 64—112.

Jurafsky, D and Manning, C (2012), 'Natural Language Processing' online course. Available at https://www.coursera.org/course/nlp

Kloptchenko, A, Magnusson, C, Back, B, Visa, A and Vanharanta, H (2004), 'Mining Textual Contents of Financial Reports', *The International Journal of Digital Accounting Research,* Vol. 4, No. 7, pages 1—29.

Lahlou, S (1996), 'A method to extract social representations from linguistic corpora', *Japanese Journal of Experimental Social Psychology,* Vol. 35, No. 3, pages 278—291.

Li, W P, Azar, P, Larochelle, D, Hill, P and Lo, A W (2015), 'Law is Code: A Software Engineering Approach to Analyzing the United States Code', *Journal of Business & Technology* Law, Vol. 10, No. 2, pages 297—374.

Loughran, T and McDonald, B (2011), 'When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks', *Journal of Finance*, Vol. 66, No. 1, pages 35—65.

Masawi, B, Bhattacharya, S and Boulter, T (2014), 'The power of words: A content analytical approach examining whether central bank speeches become financial news', *Journal of Information Science*, Vol. 40, No. 2, pages 198—210.

McLaren, N and Shanbhogue, R (2011), 'Using Internet Search Data as Economic Indicators', *Bank of England Quarterly Bulletin,* Vol. 51, No. 2.

Mihalcea, R and Tarau, P (2004), 'TextRank: Bringing order into texts', *Association for Computational Linguistics,* Proceedings of EMNLP 2004, pages 404—411.

Moniz, A and de Jong, F (2011), 'Predicting the impact of central bank communications on financial market investors' interest rate expectations', *Lecture Notes in Computer* Science, Vol. 8798, pages 144—155.

Nergues, A, Lee, J, Groenewegen, P and Hellsten, I (2014), 'The shifting discourse of the European Central Bank: Exploring Structural Space in Semantic Networks', *Signal-Image Technology and Internet-Based Systems (SITIS), 2014 Tenth International Conference*, pages 447, 455, 23—27.

Nivre, J, Hall, J and Nilsson, J (2006), 'MaltParser: A Data-Driven Parser-Generator for Dependency Parsing', *Proceedings of the fifth international conference on Language Resources and Evaluation* (LREC2006), pages 2216—2219.

Noel-Jorand, M C, Reinert, M, Bonnon, M and Therme, P (1995), 'Discourse analysis and psychological adaptation to high altitude hypoxia', *Stress Medicine*, Vol. 11, pages 27—39.

Noel-Jorand, M C, Reinert, M, Giudicelli, S, and Dassa, D (1997), 'A New Approach to Discourse Analysis in Psychiatry, applied to Schizophrenic Patient Speech', *Schizophrenia Research*, Vol. 25, pages 183—198.

Noel-Jorand, M C, Reinert, M, Giudicelli, S, and Dassa, D (2004), 'Schizophrenia: The Quest for a Minimum Sense of Identity to Ward Off Delusional Psychosis', *The Canadian Journal of Psychiatry*, Vol. 49, No. 6, pages 394—398.

Nyman, R, Gregory, D, Kapadia, S, Ormerod, P, Tuckett, D and Smith, R (2015), **'**News and narratives in financial systems: exploiting big data for systemic risk assessment', mimeo.

O'Connor, B, Balasubramanyan, R, Routledge, B R and Smith N A (2010), 'From tweets to polls: Linking text sentiment to public opinion time series', *Proceedings of the 4th International Conference on Weblogs and Social Media*, pages 122—129.

Office for National Statistics (2007), 'UK Standard Industrial Classification of Economic Activities 2007'.

Peart, J (2013), 'How do appointment processes affect the policy outputs of monetary policy committees?' Available at http://johnpeart.org/dissertation/

Reinert, M (1983), 'Une methode de classification descendante hierarchique: application a l'analyse lexicale par contexte', *Les Cahiers de l'Analyse des Donnees*, Vol. 8, No. 2, pages 187—198.

Reinert, M (1987), 'Classification descendante hiérarchique et analyse lexicale par contexte: application au corpus des poésies d'Arthur Rimbaud (Descending Hierarchical Classification and context-based lexical analysis: application to the corpus of poems by A. Rimbaud', *Bulletin de Méthodologie Sociologique*, Vol. 13, pages 53—90.

Reinert, M (1990), 'ALCESTE. Une methodologie d'analyse des donnees textuelles et une application: Aurelia de Gerard de Nerval', *Bulletin de Methodologie Sociologique*, Vol. 26, pages 24—54.

Reinert, M (1993), 'Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars. ("Lexical Worlds" and their "logics" through the statistical analysis of a corpus of narratives of nightmares)', *Langage et Société*, Vol. 66, pages 5—39.

Reinert, M (1998a), 'Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste (What is the object of a statistical analysis of discourse? Some reflections about the Alceste solution)', *Proceedings of the 4th JADT (Journées d'Analyse des Données Textuelles)*, Université de Nice JADT.

Reinert, M (1998b), *ALCESTE users' manual (English version)*, Image, Toulouse.

Reinert, M (2003), 'Le rôle de la répétition dans la représentation du sens et son approche statistique dans la méthode Alceste (The function of repetition in the representation of meaning and its statistical approach in the Alceste method)', *Semiotica*, Vol. 147, No. 1—4, pages 389—420.

Ronnqvist, S and Sarlin, P (2012), 'From Text to Bank Interrelation Maps', *Computational Intelligence for Financial Engineering & Economics*, 2104 IEEE Conference, pages 48—54.

Rosa, C and Verga, G (2006), 'On the consistency and effectiveness of central bank communication: evidence from the ECB, *European Journal of Political Economy*, Vol. 23, No. 1, pages 146—175.

Schonhardt-Bailey, C (2005), 'Measuring Ideas More Effectively: An Analysis of Bush and Kerry's National Security Speeches', *Political Science and Politics*, Vol. 38, No. 3, pages 701—711.

Schonhardt-Bailey, C (2006), *From the Corn Laws to Free Trade: Interests, Ideas and Institutions in Historical Perspective*, MIT Press, Cambridge, MA.

Schonhardt-Bailey, C (2008), 'The Congressional Debate on Partial-Birth Abortion: Constitutional Gravitas and Moral Passion', *British Journal of Political Science*, Vol. 38, pages 383—410.

Schonhardt-Bailey, C, Yager, E, and Lahlou, S (2012), 'Yes, Ronald Reagan's Rhetoric was Unique - But Statistically, How Unique?', *Presidential Studies Quarterly*, Vol. 42, No. 3, pages 482—513.

Schonhardt-Bailey, C (2013), *Deliberating American Policy: A Textual Analysis*, MIT Press, Cambridge, MA.

Siklos, P L (2013), 'The Global Financial Crisis and the Language of Central Banking: Central Bank Guidance in Good Times and in Bad', *CAMA Working Paper 58/2013.*

Tetlock, P C (2007), 'Giving Content to Investor Sentiment: The Role of Media in the Stock Market', *The Journal of Finance*, Vol. LXII, No. 3.

Upshall, M (2014), 'Text mining: Using search to provide solutions', *Business Information Review,* Vol. 31, pages 91—99.

Vallès, D W and Schonhardt-Bailey, C (2015), 'Forward Guidance as Central Bank Discourse: MPC Minutes and Speeches under King and Carney', *presented at the Political Leadership and Economic Crisis Symposium*, Yale University (February).

Warsh, K (2014), 'Transparency and the Bank of England's Monetary Policy Committee', Review by Kevin Warsh.

Weale, A, Bicquelet A and Judith, B (2012), 'Debating Abortion, Deliberative Reciprocity and Parliamentary Advocacy', *Political Studies*, Vol. 60, pages 643—667.

Further reading

Allard, J, Catenaro, M, Vidal, J-P and Wolswijk, G (2013), 'Central bank communication on fiscal policy', *European Journal of Political Economy,* Vol. 30(C), pages 1—14.

Ampofo, L, Collister, S, O'Loughlin, B and Chadwick, A (2013), 'Text Mining and Social Media: When Quantitative Meets Qualitative, and Software Meets Humans', *New Political Communication Unit Working Paper*, October 2013 (Forthcoming in Peter Halfpenny and Rob Procter (eds) Innovations in Digital Research Methods).

Armesto, M T, Hernandez-Murillo, R, Owyang, M T and Piger, J (2009), 'Measuring the Information Content of the Beige Book: A Mixed Data Sampling Approach', *Journal of Money, Credit and Banking,* Vol. 41, No. 1, pages 35—55.

Baerg, N R and Lowe, W (2015), 'Estimating Central Bank Preferences Combining Topic and Scaling Methods', *MPRA Paper 61534*, University Library of Munich, Germany.

Blasius, J and Thiessen, V (2001), 'Methodological Artifacts in Measures of Political Efficacy and Trust: A Multiple Correspondence Analysis', *Political Analysis*, Vol. 9, No. 1, pages 1—20.

Baharudin, B, Lee, L H and Khan K (2010), 'A Review of Machine Learning Algorithms for Text-Documents Classification', *Journal of Advances in Information Technology,* Vol. 1, No. 1, pages 4—20.

Bennani, H (2014), 'The art of central banks' forward guidance at the zero lower bound', *MPRA Paper 57043*, University Library of Munich, Germany.

Bennani, H and Neuenkirch, M (2015), 'The (Home) Bias of European Central Bankers: New Evidence Based on Speeches', *University of Trier Research Papers in Economics,* No. 16/14.

Benzecri, J P (1973), *L'analyse des données. Tome1: La Taxinomie. Tome 2: L'Analyse des Correspondances (Volume 1: Taxonomy; Volume 2: Correspondence Analysis)*, Dunod, Paris.

Blake, C (2011), 'Text Mining', *Annual Review of Information Science and Technology,* Vol. 45, No. 1, pages 121—155.

Blei, D M, Ng, A Y, and Jordan, M I (2003), 'Latent Dirichlet Allocation', *Journal of Machine Learning Research*, Vol. 3, pages 993—1022.

Blei, D M (2012), 'Probabilistic Topic Models', *Communications of the ACM*, Vol. 55, No. 4, pages 77—84.

Born, B, Ehrmann, M and Fratzscher, M (2010), 'Macroprudential Policy and Central Bank Communication', *BIS Papers,* No. 60.

Brugidou, M, Escoffier, C, Folch, H, Lahlou, S, Le Roux, D, Morin-Andreani, P, and Piat, G (2000), 'Les facteurs de choix et d'utilisation de logiciels d'analyse de données textuelles (parameters for choosing and using text mining software)', *Journées internationales d'analyse statistiques des données textuelles*, pages 373—380.

Chague, F D, De-Losso, R, Giovannetti, B C and Manoel, P (2013), 'Central Bank Communication Affects Long-Term Interest Rates', *Working Papers, Department of Economics*, University of São Paulo (FEA-USP), No. 2013_07.

Drury, B (2013), 'A Text Mining System for Evaluating the Stock Market's Response to News', PhD Thesis, University of Porto.

Fisher, I E, Garnsey, M R, Goel, S and Tam, K (2010), 'The Role of Text Analytics and Information Retrieval in the Accounting Domain', *Journal of Emerging Technologies in Accounting,* Vol. 7, pages 1—24.

Geraats, P M (2010), 'Talking Numbers: Central Bank Communications on Monetary Policy and Financial Stability', Paper presented at the *5th ECB Statistics Conference "Central Bank Statistics: What Did the Financial Crisis Change?"*, 19-20 October 2010.

Greenacre, M and Underhill, L G (1982), *Scaling a data matrix in low-dimensional Euclidean space. Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge.

Greenacre, M (1984), *Theory and Applications of Correspondence Analysis*, Academic Press, London.

Greenacre, M and Hastie, T (1987), 'The Geometric Interpretation of Correspondence Analysis', *Journal of the American Statistical Association*, Vol. 82, No. 398, pages 437—447.

Greenacre, M (1993), *Correspondence Analysis in Practice*, Academic Press, London.

Grimalid, M B (2010), 'Detecting and interpreting financial stress in the Euro Area', *European Central Bank Working Paper,* No. 1214.

Grimmer, J (2010), 'A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases', *Political Analysis,* Vol. 18, No. 1, pages 1—35.

Gupta, R and Gill, N S (2012), 'Financial Statement Fraud Detection using Text Mining', *International Journal of Advanced Computer Science and Applications,* Vol. 3, No. 12, pages 189—191.

Gupta, V and Lehal, G (2009), 'A Survey of Text Mining Techniques and Applications', *Journal of Emerging Technologies in Web Intelligence,* Vol. 1, No. 1, pages 60—76.

Holmstrom, B (1999), 'Managerial Incentive Problems: A Dynamic Perspective', *Review of Economic Studies*, Vol. 66, No. 1, pages 169—82.

Hotho, A, Nürnberger, A and Paaß, G (2005), 'A brief survey of text mining', *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, Vol. 20, No. 1, pages 19—62.

Jurafsky, D and Martin, J H (2008), *Speech and Language Processing 2nd ed.,* Prentice Hall.

Kearney, C and Liu, S (2014), 'Textual sentiment in finance: A survey of methods and models', *International Review of Financial Analysis,* Vol. 33, pages 171—185.

Kent, E (2014), 'Text Analytics—techniques, language and opportunity', *Business Information Review*, Vol. 31, No. 1, pages 50—53.

Lewis, D D, Yang, Y, Rose, T G and Li, F (2004), 'RCV1: A New Benchmark Collection for Text Categorization Research', *Journal of Machine Learning Research,* Vol. 5, pages 361—397.

Manning, C D, Raghavan, P and Schutze, H (2008), *Introduction to Information Retrieval*, Cambridge University Press.

Meade, E and Stasavage, D (2008), 'Publicity of Debate and the Incentive to Dissent: Evidence from the US Federal Reserve', *Economic Journal,* Vol. 118, No. 528, pages 695—717.

Montoyo, A, Martínez-Barco, P and Balahur, A (2012), 'Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments', *Decision Support Systems,* Vol. 53, No. 4, pages 675—679.

Nadkarni, P M, Ohno-Machado, L and Chapman, W (2011), 'Natural language processing: an introduction', *Journal of the American Medical Informatics Association,* Vol. 18, No. 5, pages 544—551.

Nagler, J and Tucker, J A (2015), 'Drawing Inferences and Testing Theories with Big Data', *Political Science & Politics,* Vol. 48, No. 1, pages 84—88.

Nassirtoussi, A K, Aghabozorgi, S, Wah, T Y, Chek, D and Ngo, L (2014), 'Text mining for market prediction: A systematic review', *Expert Systems with Applications,* Vol. 41, No. 16, pages 7653—7670.

Nergues, A, Hellsten, I and Groenewegen, P (2015), 'A Toxic Crisis: Metaphorizing the Financial Crisis', *International Journal of Communication,* Vol. 9, pages 106—132.

Prat, A (2005), 'The Wrong Kind of Transparency', *American Economic Review*, Vol. 95, No. 3, pages 862—877.

Quinn, K M, Monroe, B L, Colaresi, M, Crespin, M H and Radev, D R (2010), 'How to Analyze Political Attention with Minimal Assumptions and Costs', *American Journal of Political Science,* Vol. 54, No. 1, pages 209—228.

Sarda, V, Sakaria, P and Mistry, D (2014), 'Fraud Detection in Financial Statements Using Classification Algorithm', *International Journal of Emerging Technology and Advanced Engineering',* Vol. 4, No. 9, pages 561—563.

Schumaker, R P and Chen, H (2009), 'Textual Analysis of Stock Market Prediction Using Financial News: The AZFin text system', *ACM Transactions on Information Systems,* Vol. 27, No. 2, Article No. 12.

Steinbach, M, Karypis G and Kumar V (2000), 'A Comparison of Document Clustering Techniques', University of Minnesota *Technical Report,* No. 00-034.

Taffler, R J and Tuckett, D (2010), *Emotional finance: the role of unconscious in financial decisions*, Behavioural finance: Investors, corporations and markets (eds H. K. Baker and J. R. Nofsinger), John Wiley & Sons, Inc., Hoboken, NJ, USA.

Tobback, E, Daelemansb, W, Fortunya, E J, Naudts, H and Martensa, D (2014), 'Belgian Economic Policy Uncertainty Index : Improvement through text mining', In *ECB Workshop*, 7-8 April 2014.

Tuckett, D, Nyman, R, Ormerod, P and Smith, R (2014), 'Big Data and Economic Forecasting: A Top-Down Approach

Using Directed Algorithmic Text Analysis', *ECB Workshop on Big Data for Forecasting and Statistics*.

Tuckett, D, Smith, R E, and Nyman, R (2014), 'Tracking phantastic objects: A computer algorithmic investigation of narrative evolution in unstructured data sources', *Social Networks,* Vol. 38, pages 121—133.

Weller, S C and Romney, A K (1990), *Metric Scaling: Correspondence Analysis*, Sage Publications, London.

Glossary

Abduction Reasoning method that attempts to infer the best explanation for a particular event based on some data, without ambition to generate an explanation generalisable to other cases.

Alceste Textual data analysis software; acronym for *Analyse des Lexèmes Co-occurents dans les Énnoncés Simples d'un Texte* (Analysis of the Co-occurring Lexemes within Simple Statements of Text).

Bag-of-words representation Representation of documents as 'bags' of words that accounts for their frequencies and disregards word order and grammar.

Bayes' rule In text mining, Bayes rule states that the most likely class for a document is the class that maximises the product of the prior probability of the class and the likelihood of words assigned to the class.

Bayesian Information Criteria (BIC) Method to measure models' fitness, penalising for the number of estimated parameters.

Big data Data displaying one or more of the following characteristics: high volume, high velocity and/or qualitatively various.

Boolean search Type of search that combines keywords with logical operators such as AND, NOT and OR.

Case folding Process of converting all alphabetic characters to lowercase.

Chi-square (χ^2) value Statistical measure of significant difference between observed and expected frequencies.

Content analysis Processes and techniques used to detect topics and/or sentiment in text, based on the frequencies of the words.

Co-occurrence Occurrence of two terms in a corpus alongside each other.

Corpus Collection of documents (plural: corpora).

Correspondence Analysis Statistical technique aimed at obtaining a lower dimensional graphical representation of the data. The positions of the points is contingent on correlations and the distance between points reflects the degree of co-occurrence.

Cosine similarity Measure of the angle formed by two vectors.

Covariate Explanatory or independent variable.

Cyclomatic complexity In text mining, measures the complexity derived from the number of conditional statements used in the text, e.g. 'if', 'except', 'in the event'.

Deduction Reasoning method that starts from a general theory and then uses particular datasets to test its validity.

Descending Hierarchical Classification Partitioning algorithm in Alceste used to identify statistically significant relationships between words and elementary context units. The algorithm seeks to maximise the χ^2 values of a contingency table, and tests if further splits improves the χ^2 values. The iterative process comes to conclusion when a predetermined number of iterations no longer results in statistically significant divisions.

Dimensionality Number of words (in text mining) or variables.

Disambiguation Process of finding the most probable meaning of a word in a specific phrase.

Duplication The fact that there may be more than one instance of the same observation (document) in a sample (corpus).

Elementary Context Units In Alceste, Elementary Context Units ("ECUs") or statements are constructed based on the punctuation in the text.

Entropy Measure of dispersion in data (see also Shannon entropy).

Euclidean distance Distance between points defined by the square root of the sum of the squared differences between the respective coordinates of the points.

Hypernym Word which meaningfully encompasses more specific words, e.g. 'financial institution' is a hypernym of 'bank' and 'credit union'.

Hyponym Word whose meaning is included in the meaning of another more general word, e.g. 'bank' and 'credit union' are hyponyms of 'financial institution'.

Induction Reasoning method that starts from data without priors and seeks to generate *general* theoretical claims.

Information Content Negative log of the probability of a word in the text.

Laplace smoothing Simple smoothing technique consisting of adding a number—e.g. 1—to every word count so no word has a zero probability estimate.

Latent Dirichlet Allocation (LDA) Mixed-membership model representing topics as probability distributions over the terms in the vocabulary.

Latent Semantic Analysis (LSA) Latent variable model representing the document-term matrix in terms of its singular value decomposition.

Latent variable models Models conceptualising texts as containing latent variables, and then empirically deriving these latent variables on the basis of observed words.

Least common subsumer Most specific word that is a shared ancestor of two words.

Lemmatisation Technique using part-of-speech tagging to classify each word as a (word, part of speech) pair for converting the word into a base form, e.g., 'saw' tagged as a verb would become 'see', but not when tagged as a noun.

Lexicon Vocabulary of a particular language.

Likelihood In text mining: probability of words given a class based on a prior observed distribution.

Mixed-membership model Latent variable model in which each document can belong to multiple topics.

Naïve Bayes classifier Simple classification algorithm based on Bayes' rule.

Natural language processing (or computational linguistics) Set of techniques for computational processing and analysis of naturally occurring human languages.

Non-Negative Matrix Factorization Decomposition technique used to approximate multivariate data by non-negative factors.

Normalised count Count in a class relative to the total number of observations.

Optical character recognition software Software that converts printed text into machine-readable text.

Orthogonal Two vectors are orthogonal if they intersect at a right angle.

Over-fitting The occurrence of patterns in the training data that do not persist in unseen data.

Parsing algorithm Statistical algorithm used to detect dependencies and syntax in strings of text.

Part of speech (POS) tagging Form of linguistic analysis consisting of identifying the grammatical class to which each word belongs.

Path-length similarity Distance measure capturing the similarity of terms in a hypernym hierarchy.

Phi (ϕ) coefficient Measure of association between two binary variables.

Polarity detection Binary classification of positive and negative sentiments.

Polysemy The fact that the same word can have multiple meanings in different contexts, e.g. 'tank' can mean a fish aquarium or a military vehicle.

Principal components Set of variables determined by Principal Component Analysis that attempt to capture most of the variance in the data.

Prior probability Probability of a class occurring in the training data.

Regular expression Sequence of characters used to search for patterns in text and split the text on these patterns. For example, a search for the words 'Bank of England' might use a regular expression that searches for a word that begins with the letter 'B' followed by: three alphabetical characters, a space, the word 'of', a space and a word that begins with the letter 'E'.

Resnik measure Similarity measure taking as input two words. The technique uses the idea of a least common subsumer.

Semantics Meaning of text.

Sentiment analysis (or opinion mining/subjectivity analysis) Techniques aimed at detecting sentiment about a subject.

Shannon entropy Measure of dispersion in a random variable.

Single-membership model Topic model in which each document can belong to only one topic.

Singular Value Decomposition (SVD) Technique providing the linear combinations of terms that explain most of the variance of terms across documents, as well as the linear combinations of documents that explain most of the variance of documents across terms.

Stem Part of a word after stemming is applied, or part of a word to which affixes can be attached, e.g. the word 'banking' contains the stem 'bank' and the affix '-ing'.

Stemming Technique that consists of using a deterministic rule for removing word endings.

Stop-words Very common words that contribute little to distinguishing the content of one document from the content of another, e.g. 'the', 'of'.

Structured data Data stored in fields in a traditional database.

Supervised machine learning Set of techniques that learn from *classified* observations and then assign classes on unseen data, based on the prior observed distribution.

Synonymy Characteristic of words sharing the same underlying concept.

Target Variable being predicted by a supervised machine learning algorithm.

Term frequency−inverse document frequency (tf.idf) Common weighting scheme used in the text mining literature that consists of giving lesser weight to words that appear more frequently and greater weight to words that appear less frequently.

Term-document matrix Matrix capturing the frequency of each term in the set of documents.

Test data Set of data used for validation of the 'trained' algorithm, e.g. to detect over-fitting.

Text mining Umbrella term for a range of computational tools and statistical techniques that quantify text.

Thesaurus-based algorithms Similarity algorithms using online thesaurus.

Tokens Words, phrases, or symbols in text.

Training data In text mining, set of classified data used to 'train' a supervised algorithm to 'learn' the co-occurrence of classes and words.

Unstructured data Data not residing in a traditional row/column database format.

Unsupervised machine learning Set of techniques that involve taking *unclassified* observations (text or otherwise) and uncovering hidden patterns.

Vector space model Unsupervised machine learning model that represents documents as vectors of word frequencies and measures similarity using a specified distance function.

VIX index Measure of market expectations of option price's volatility on the S&P 500 stock index.

Word cloud Pictorial cluster of the frequencies of discrete words, where the size of each word indicates its frequency.

X-means clustering algorithm Extension of k-means clustering algorithm, which automatically learns the optimal number of clusters.

Zipf's law Observed empirical regularity that the frequency of a word is inversely proportional to its frequency rank.