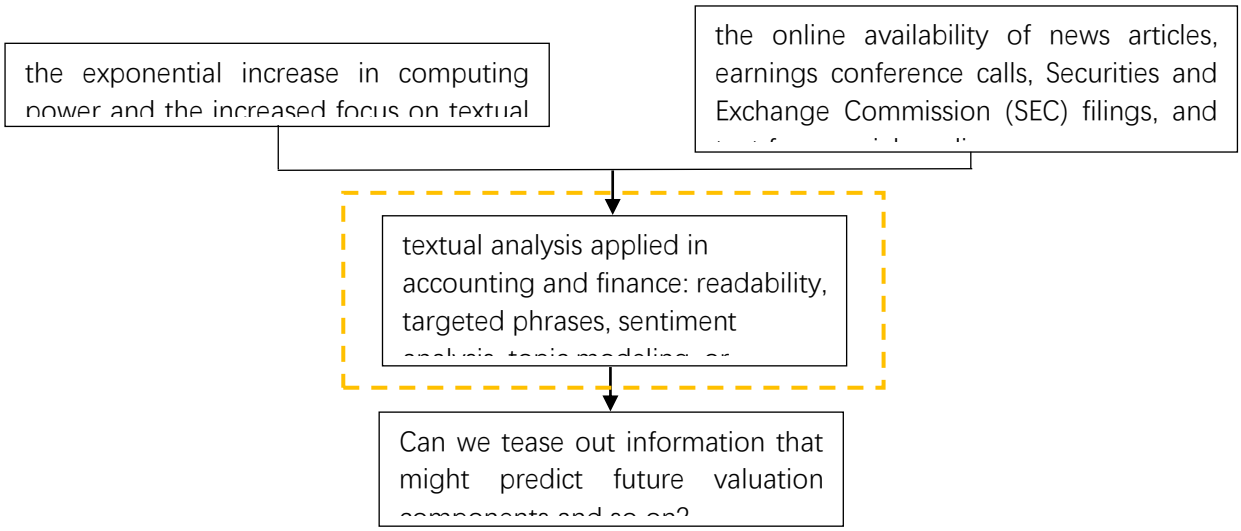
Textual Analysis in Accounting and Finance: A Survey

Tim Loughran, Bill Mcdonald

Journal of Accounting Research, 2016

1. Introduction

1.1 Background



1.2 Main Work

We fold a more selective, focused survey of the accounting, finance, and economics literature on textual analysis into a description of some of its methods, and emphasize some points such as the importance of exposition, transparency and replicability.

2. Information Content, Document Structure, and Readability

We begin with a discussion of the fundamental issue of information extraction in textual analysis and document structure. Document structure is reflected by the graphic design along with the writing style (readability) used to convey the information.

2.1 INFORMATION CONTENT

Textual analysis is most notably demarcated from quantitative analysis by its imprecision.

- The meaning of the characters is not unambiguous and in most cases depends substantively on the context.
- For many SEC filings prior to roughly 2005, there is a lack of consistency in HTML formatting.
- If best is a positive word and the document is not parsed to exclude company names, firms such as Best Buy will have very positive sentiment measures. Astonishing seasonality can be caused by may if it is included as a measure of uncertainty.

Sum: The ideal hypotheses in textual analysis are the ones based on straightforward characteristics of the data. We must be careful that the imprecision of the method does not overwhelm any hoped for gains in identifying meaning.

2.2 READABILITY

We consider the overarching issue of whether the receiver of information can accurately reconstruct the intended message.

2.2.1. Examples of Studies Using Readability

- Fog Index = 0.4(average number of words per sentence + percentage of complex words); estimates the number of years of education needed to understand the text on a first reading (Lewis et al. [1986]).
- The first paper to examine the link between 10-Ks readability and firm performance for a meaningful sample is Li [2008], finding that firms with lower reported earnings tend to have annual reports that are harder to read (poorly performing firms needing to have more text and longer sentences to fully explain their situation to investors), vice versa.
- Li [2008], Lundholm, Rogo, and Zhang [2014], Lehavy, Li, and Merkley [2011], De Franco et al. [2015], Rennekamp

[2012]

2.2.2. Defining and Measuring Readability

Many have questioned Fog Index usage for business documents.

- Loughran and McDonald [2014] empirically demonstrate that the Fog Index is a poorly specified readability measure when applied to business documents because the vast majority of these documents are not distinguished by writing style. They show that syllable counts are a poor measure of readability for business documents.
  New method: natural log of gross 10-K file size (reflect the underlying complexity of the firm's business). They find that firms with bigger 10-K file sizes are significantly linked with larger subsequent stock return volatility and so on.
- Log file size is not a perfect measure of 10-K readability because firms responded to the Enron accounting scandal by expanding the number of pages in their annual report to improve their firm-specific transparency.

Sum: The use of readability measures must consider the context of application.

3. Bag-of-Words Methods and the Term-Document Matrix

We now focus on the methods attempt to computationally distill meaning from the message.

Bag-of-words: Methods where we assume that the order, and thus direct context, of a word is unimportant. Many of these are based on collapsing a document down to **a term-document matrix** consisting of rows of words and columns of word counts, allows the computational task of summarizing a large document to be simplified by orders of magnitude.

3.1. TARGETED PHRASES

The researcher target a few specific words or phrases. Because of ambiguity, large word lists are much more prone to error when compared to tests focusing on a few unambiguous words or phrases.

- Loughran, McDonald, and Yun [2009] consider the frequency of the word "ethic" (and its variants) along with the phrases "corporate responsibility," "social responsibility,"and "socially responsible" in 10-K filings to determine if these counts are associated with "sin" stocks, corporate governance measures, and class action lawsuits.

3.2 WORD LISTS

Compiling word lists that share common sentiments, researcher can count words associated with each attribute and provide a measure of sentiment.

- Feature:
  - avoid researcher subjectivity
  - the method can be scaled to large samples
  - straightforward to replicate the analysis
  - It faces the homographs challenge and discards information that can be distilled from word sequences.

3.2.1. The Henry [2008] Word List

The first word list we are aware of that was created for financial text.

- Advantages: created by examining earnings press releases for the telecommunications and computer services industries
- Disadvantages: the limited number of words contained in the list especially for negative words
- Application:
  - Price et al. [2012] use this lists to gauge tone during quarterly earnings conference calls for publicly traded stocks and assert that the Henry [2008] dictionaries are better at measuring the tone of earnings conference calls than the Harvard IV-4 word lists.
  - Doran, Peterson, and Price [2012], Davis et al. [2015]

3.2.2. Harvard GI Word Lists

The first ones readily available.

- Application:
  - Tetlock [2007] links the tone of the WSJ's "Abreast of the Market" daily column with stock market levels.
  - Tetlock, Saar-Tsechansky, and MacSkassy [2008] examine WSJ and Dow Jones News Service stories on S&P 500 firms.

- Kothari, Li, and Short [2009] examine the relation between the content of disclosures by firms, analysts, and news outlets and stock return volatility as well as analyst forecast error dispersion.

3.2.3. Diction Optimism and Pessimism Word Lists

It has 35 different dictionary subcategories.

- Application:
  - Davis, Piger, and Sedor [2012] find that firms with more positive tone in their earnings press releases are associated with higher subsequent ROA.
  - Rogers, Van Buskirk, and Zechman [2011] examine the relation between Diction net tone and shareholder litigation.

3.2.4. Limitations of the Harvard and Diction Sentiment Word Lists

Li [2010b] finds no positive relation between the tone of the MD&A section of the 10-K using the GI and Diction dictionaries and future performance and report that almost the vast majority of the Diction words and the Harvard GI negative words do not have pessimistic meaning when used in the context of financial documents such as tax, cost, capital, board, liability, depreciation, crude, cancer, and mine, vice versa.

Sum: Word lists designed specifically for business communication should be used to measure the sentiment of business text.

3.2.5. Loughran and McDonald [2011] Word Lists

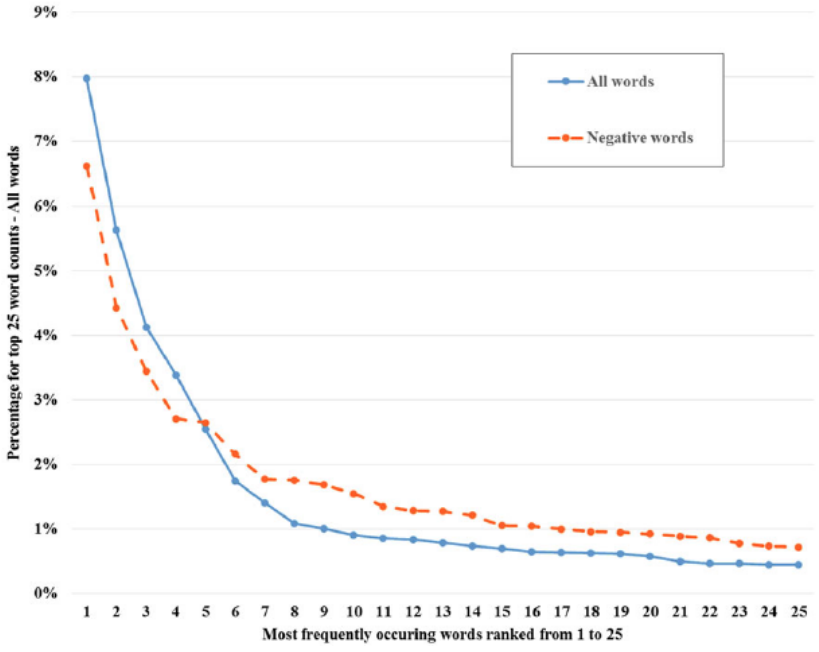
Loughran and McDonald [2011] created six different word lists (negative, positive, uncertainty, litigious, strong modal, and weak modal) by examining word usage in a large sample of 10-Ks during the period 1994–2008.

- Advantages: It is comprehensive and created with financial communication in mind.
- Application: the L&M lists have become predominant in more recent studies
  - Feldman et al. [2010] use it to examine the market's immediate response to changes in MD&A tone.
  - Mayew and Venkatachalam [2012] finds that manager's voice during their discussion with analysts in a conference call are associated with returns.
  - Many papers have used the it to measure tone in newspaper columns, such as Dougal et al. [2012], Garcia [2013], Gurun and Butler [2012], or in news articles such as Liu and McConnell [2013].

3.2.6. Zipf's Law

Word counts tend to follow a power law distribution and certain words can potentially have a large impact on the results.

Sum: Research using word classifications must identify the proportions of the most frequently occurring words so that the reader can determine if misclassification is driving the paper's results (vice, board, liability, tire, and depreciation).



3.2.7. Term Weighting

How these counts should be normalized? Proportions? In some instances, we might also want to adjust a word's weight in the analysis based on how unusual the term is. Perhaps the more unusual words should receive more weight?

- Loughran and McDonald [2011] consider one of the more common term weighting schemes from the literature labeled *tf–idf* (term frequency inverse document frequency), and find that this approach produces regressions with better fit than the approaches using simple proportions.
  - The inverse document frequency is: $idf_t = log(N/df_t)$, when df_t as the number of documents in a collection of documents containing the term t and N represent the total number of documents in the collection.
  - Note that $tf_{t,d}$ is the raw count of term t in document d, and a_d is the average word count in document d, then

$$tf\text{-}idf_{t,d} = \begin{cases} \dfrac{(1 + \log(tf_{t,d}))}{(1 + \log(a_d))} log\dfrac{N}{df_t} & if\ tf_{t,d} \geq 1 \\ 0 & otherwise \end{cases}$$

3.3 NAIVE BAYES METHODS

It is the most popular approaches for word classification using supervised machine learning.

- Advantages:
  - it is one of the oldest, most established methodologies to analyze text
  - large corpuses of data can easily be included in the analysis
  - once the rules/filters of gauging the text are established, no additional researcher subjectivity affects the measuring
- Disadvantages:
  - difficult to replicate since the procedure has literally hundreds of various unpublished rules to measure the context
  - need for the researcher to fully reveal the words driving the empirical classifications
- Applications:
  - The earliest use in finance is Antweiler and Frank [2004]. They examine 1.5 million stock message postings on Yahoo! Finance and Raging Bull for a small number of firms and find that stock message board postings are related to stock market levels, trading volume, and volatility.
  - Li [2010b]/Huang, Zang, and Zheng [2014] uses the Naive Bayes method to examine the content of forward-looking statements (FLS) in the MD&A section of the 10-K/analyst reports.

Sum: Which method produces the most discerning classification of sentiment is yet to be determined.

3.4. THEMATIC STRUCTURE IN DOCUMENTS

These techniques can be used to classify common themes in documents. One of the earliest approaches is latent semantic analysis (LSA) whose distinguishing feature is that it can avoid the limitations of count-based methods associated with synonyms and polysemy (terms with multiple meanings).

- Applications: LSA was first used in business by Boukus and Rosenberg [2006], who analyze the information content of the Federal Open Market Committee's minutes.
- Evolution:
  - probabilistic LSA (pLSA): based on a latent class model (see Hofmann [2001]), uses singular value decomposition to identify an orthogonal basis within the dimensionality constraint
  - latent Dirichlet allocation: Dirichlet-based priors (LDA, see Blei, Ng, and Jordan [2003]), uses a Bayesian model that views the documents as a mixture of latent topics

4. Document Narrative

Although word choice is important, the essential character of any text is based on how the story is told through the sequencing
of words. We suggest the following hierarchy of analysis: lexical, collocation, syntactic, semantic, pragmatic, and discourse.

- Lexical: parsing the document's characters into chunks of words or meaningful tokens
- Collocation: meaning of words can be derived from their collocation with other words
- Syntactic: derive additional information by examining the grammatical structure of the sentence

- Semantics: infer meaning within the context of the sentence
- Pragmatics: infers meaning with context provided by external knowledge
- Discourse: derive meaning from the collective document

Sum: Thus far, applications in accounting and finance are predominately in the initial phase.

- Application: Allee and DeAngelis [2015] provide a good example in accounting and finance of a first step beyond the bag-of-words approach by measuring tone dispersion.

5. Measuring Document Similarity

- Two documents d_1 and d_2 that have been collapsed into two vectors x and y of word counts, the cosine similarity measure for $i = 1$ to N words is defined as

$$\text{cosine similarity}\,(d_1, d_2) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

- Application: Brown and Tucker [2011], Hoberg and Phillips [2015] and Lang and Stice-Lawrence [2015]

Egozi, Markovitch, and Gabrilovich [2011] propose an interesting combination of cosine similarity and LSA.

6. Implementation: Tripwires, Technology, and a Simple Example

6.1 WHAT IS A WORD?

Parse each document into a vector of tokens with the help of word lists. Moreover, Loughran and McDonald [2011] develop a word list for business-related textual analysis that is based on the "2of12inf" dictionary.

6.2 WHAT IS A SENTENCE?

In many studies focusing on readability, the researcher is required to calculate the average number of words per sentence. A typical attempt will first remove abbreviations, headings, and numbers (with decimals).

6.3 WHY POSITIVE TONE OR NET TONE IS PROBLEMATIC

Negative words seem unambiguous. Positive words in addition to their positive usage, are just as frequently used to frame a negative statement. Consider a simple case: A careful manager might use 90% positive words in dismissing an employee.

6.4 TARGETING SECTIONS IN MANDATED DISCLOSURES

Focus your analysis on a specific section of the document. An additional problem for researchers focusing on only one section is that companies can shift content between sections.

6.5 LEVELS VERSUS DIFFERENCES

Some of the issues caused by misclassification can be mitigated in word-count methods by differencing.

Note that differencing would imply a scenario where the reader was making a year-to-year comparison of tone.

6.6 HOW TO IMPLEMENT

6.6.1. Programming Languages for Textual Analysis

- Popular language: Perl → Python
- Key component: regular expression
- Note that using prepackaged programs to parse business-related documents may creates significant uncertainty about the accuracy. However, a norm where each researcher produces independent code does not seem efficient. So that we aim to establish a repository with standardize routines used in textual analysis in accounting and finance.

6.6.2. A Simple Example

This table presents regressions that test the impact of using the term "non-GAAP" in 10-K filings. The sample and control variables are defined in Loughran and McDonald [2014].

TABLE 1

Use of "Non-GAAP" in 10-K Filings and Market Model Postfiling Date Root Mean Square Error

Dependent Variable = Postfiling Date RMSE	(i)	(ii)
Non-GAAP dummy		0.054***
		(2.64)
Log (file size)	0.073***	0.072***
	(4.60)	(4.57)
Prefiling market model alpha	−0.898***	−0.898***
	(−4.06)	(−4.06)
Prefiling market model RMSE	0.536***	0.535***
	(11.89)	(11.90)
Abs (filing period return)	5.046***	5.048***
	(17.56)	(17.59)
Log (size in $ millions)	−0.117***	−0.118***
	(−5.91)	(−5.87)
Log (book to market)	−0.140***	−0.140***
	(−2.52)	(−2.53)
NASDAQ dummy	0.264***	0.265***
	(3.38)	(3.38)
Intercept	1.537***	1.491***
	(6.29)	(5.96)
Industry dummies	Included	Included
Year dummies	Included	Included
No. of observations	66,707	66,707
Adj. R^2	46.96%	46.97%

7. Areas for Future Research in Textual Analysis

- disentangling the role of firm-level complexity from readability → focus on broader concept of information complexity
- term weighting lack theoretical motivation, provides too many degrees of freedom → a structured analysis
- simple positive/negative dichotomy of sentiment analysis → other word groupings might produce useful targets
- potential words might impact sentiment measures within the context of different bodies of text → transparently modifying the word lists to fit specific application
- parsing social media is even more challenging → develop methods that are better able to capture the information
- we have only focused on textual analysis in the English language → linguistic typologists can provide useful insights

8. Conclusion

- The traditional concept of readability does not map well into determining the effectiveness of business documents as information conduits.
- Zipf's law documents the fact that a very small number of words will dominate the frequency counts.
- Avoid using word lists and algorithms derived in the context of other disciplines unless they are proven to be effective.
- Parsing methods must be documented in detail, especially for cases where the underlying documents have complex structures.
- Provide a description of your method that makes your research replicable.

金融市场文本情绪研究进展

唐国豪，姜富伟，张定胜
《经济学动态》2016 年第 11 期

一、引言

二、情绪影响资产价格的理论基础与早期研究

（一）情绪影响资产价格的理论基础

1、前景理论、过度自信和保守性偏差 → 投资决策不是完全理性而是存在情绪化倾向的

2、DSSW 噪音交易者模型 → 由于市场上一部分噪音交易者受到非理性情绪影响，当情绪过度乐观或悲观时，相对于理性投资者来说，其对风险资产的需求往往过度旺盛或不足，进而对风险资产的价格产生影响。

（二）情绪测度的早期实证研究

1、通过直接调查法（电话访谈、问卷调查等方式）来衡量金融市场情绪

2、使用外生非金融市场变量作为衡量市场情绪的指标，进而预测股票市场收益

3、采用与金融市场交易相关的一个或多个变量作为衡量情绪的指标，构建单一性指标或者综合性指数。目前广泛接受的是由封闭式基金折价、换手率等六个代理变量组成的运用主成分分析法构建的 BW 指数

三、提取和分析文本情绪的主要方法与研究进展

（一）词汇分类字典法

（二）文本词汇加权

（三）基于机器学习的朴素贝叶斯法

（四）基于其他文本特征的分析

1、文本可读性

2、文本叙述方法

3、文本主题结构和相似性

四、文本情绪与资产预期回报

（一）横截面实证检验研究

1、投资组合分析法

2、法玛—麦克白回归法

（二）时间序列实证检验研究

采用线性预测性回归模型检验情绪与市场整体回报间关系，主要方法包括样本内预测和样本外预测法。

$$R_{OS}^2 = 1 - \frac{\sum_{t=p}^{T-1}(R_{t+1}^m - \hat{R}_{t+1}^m)^2}{\sum_{t=p}^{T-1}(R_{t+1}^m - \bar{R}_{t+1}^m)^2}$$

通过比较上述统计量的大小，可以检验时间序列上不同情绪指数对股票市场整体收益预测能力的大小。