

International Conference on Business and Economics - Hellenic Open University

Vol 5, No 1 (2025)

Proceedings of the ICBE-HOU 2025



Textual analysis of sentiment in finance and accounting: A comprehensive review

Virginia Sotiropoulou, Vasilios Sogiakas

Copyright © 2026, International Conference on Business and Economics - Hellenic Open University



This work is licensed under a [Creative Commons Attribution-NonCommercial 4.0](https://creativecommons.org/licenses/by-nc/4.0/).

To cite this article:

Sotiropoulou, V., & Sogiakas, V. (2026). Textual analysis of sentiment in finance and accounting: A comprehensive review. *International Conference on Business and Economics - Hellenic Open University*, 5(1). Retrieved from <https://eproceedings.epublishing.ekt.gr/index.php/ICBE-HOU/article/view/9998>

Textual analysis of sentiment in finance and accounting: A comprehensive review

Virginia Sotiropoulou^{*1} and Vasilios Sogiakas²

¹Postdoctoral researcher, Accounting Lab, School of Social Sciences, Hellenic Open University, Greece.

University Campus, 26335, Patras, Greece.

E-mail: virginia.sotiropoulou@ac.eap.gr

²Professor, School of Social Sciences, Hellenic Open University, Greece.

University Campus, 26335, Patras, Greece.

E-mail: vsogiakas@eap.gr

Abstract

The purpose of this study is to examine critically the literature of textual analysis of disclosures' sentiment in finance and accounting. Specifically, this research work investigates, compares, and contrasts the various methodologies that have been implemented to date to measure textual sentiment of financial and accounting information. Furthermore, we present the attributes of textual information of disclosures by investigating thoroughly established research in the field. We summarize the significant and influential empirical findings regarding textual analysis of sentiment in finance and accounting. Moreover, emerging opportunities that hold promise for future research are appearing from the incorporation of artificial intelligence techniques and the extensive use of computer linguistic methods to extract the sentiment in finance and accounting research. The adoption of these methodologies for analyzing textual information in finance and accounting literature could enable researchers to evaluate more efficiently the quality of financial reporting.

JEL Classifications: G14, M41

Keywords: Finance, accounting, textual analysis, sentiment

Funding: The research work was supported by the Hellenic Open University (HOU) (Code of project: 80687).

* (Corresponding author) Hellenic Open University, School of Social Sciences, virginia.sotiropoulou@ac.eap.gr

1 Introduction

Prior literature indicates that information plays a crucial role in the documentation of a firm's operations (Loughran & McDonald, 2016). Initially, Frazier et al. (1984), point out the fact that narrative information of accounting disclosures is ignored most of the times. Furthermore, they suggest combining narratives along with ratios of firm performance and results of operations by using a “computer analysis system.” Iker (1974) is the first researcher that introduces the notion of the computer analysis system for textual analysis of the sentiment of narratives.

Loughran and McDonald (2016) indicates the contribution of the increased computer power during the last decades to the evolution of textual analysis methods. Specifically, they state that textual analysis has gained prominence as a developing field within finance and accounting. The focus on this methodology has been increased especially after the scandals in the USA and Europe as the attention to the quality of financial reports has been augmented according to DeZoort et al. (2002) (Al-Shaer et al., 2017). Therefore, the accuracy and transparency of financial reporting has major importance in the post-Enron era (Al-Saer et al., 2017). The establishment of regulations such as the Sarbanes-Oxley Act of 2002 and the XBRL mandate of 2009 indicated the significance of the reliability and quality of financial reporting.

Therefore, after the development of textual analysis methods, academics expanded their research by using not only the ratios from financial disclosures, but also textual analysis variables in the same research model to capture the sentiment of the reports: Loughran and McDonald (2014a), Loughran and McDonald (2014b), Lawrence (2013), Kravet and Muslu (2013), Li and Zhao (2015).

Textual analysis has also been extensively used in accounting disclosures, as well: Loughran and McDonald (2011); (2016), Al-Shaer et al. (2017), and Eachempati et al. (2021).

Finally, textual analysis has been used to analyze corporate disclosures. The pioneering paper that introduced “content analysis” of corporate disclosures was Ingram and Frazier’s: “Environmental Performance and Corporate Disclosure”, in 1980. Moreover, prior literature has investigated sustainability (Laine, 2010), environmental (Nazari et al., 2017), and corporate social responsibility disclosures (Muslu et al., 2019). Finally, recent research has examined the use of artificial intelligence in analyzing corporate disclosures (Cao et al., 2023).

In this paper, we review the rapidly growing body of research that implements textual analysis to capture the sentiment of disclosures in the fields of accounting and finance. We present thoroughly the predominant methodologies including dictionary-based approaches, machine learning techniques as well as the programming languages that are implemented to define the sentiment of textual information. Furthermore, we outline the key characteristics

of textual information in finance and accounting based on the findings of established literature in these areas. Next, we describe and discuss the main empirical results of significant research works regarding textual analysis of sentiment in finance and accounting to date. We also refer to the methodology that is adopted in each study. Finally, we attempt to highlight several areas for further research.

The remainder of our paper is structured as follows. Section 2 describes and discusses the methodologies implemented to extract textual sentiment of disclosures in finance and accounting. Section 3 introduces the attributes of financial and accounting narratives according to distinguished literature in the field. Section 4 presents and reviews high-impact research regarding textual analysis that captures the sentiment of disclosures in finance literature, while empirical research in textual analysis of accounting information is described in detail in Section 5. Section 6 outlines potential areas for future study. Finally, Section 7 summarizes the present work and draws together the most significant conclusions.

2 Methodologies for measuring textual sentiment

Textual information of financial and accounting disclosures requires a unique processing, as it must be transformed first to quantitative data and then it can be analyzed. In specific, to analyze financial ratios along with textual variables, it is necessary to quantify first the information content of reports. Therefore, analyzing text can be divided in three steps: collecting text, cleaning text to have an appropriate format, and finally analyzing text by using various methods (Guo et al., 2016).

2.1 Dictionaries

Dictionaries use word lists to classify words of financial and accounting disclosures in the appropriate category and extract the meaning of the text based on its tone. According to Loughran and McDonald (2016) there are four word lists that can be used to investigate the sentiment of the text: Harvard's GI, Henry, Diction, and Loughran and McDonald dictionary (2011). The main drawback of Harvard's GI is that it is not designed exclusively for business context, which means that the words are often misclassified. The same phenomenon is observed in the case of Diction. Hence, Loughran and McDonald (2016) reach the conclusion that when textual analysis is implemented in financial or accounting texts, the method chosen must be the most appropriate and suitable for the context of the text to interpret correctly the sentiment of the disclosures.

Loughran and McDonald examine the effectiveness of dictionaries in textual analysis of the sentiment of financial and accounting disclosures and they find that words with unfavorable

meaning that are suitable for other scientific fields do not fit in with finance and accounting context. They create their own negative word list specially adjusted to business content. They also create five alternative word lists to capture the tone of the text (“positive, uncertainty, litigious, strong modal, and weak modal word lists”) (Loughran & McDonald, 2011).

Furthermore, Matthies and Coners (2015) examine a combination of dictionaries and statistical techniques. Their research indicates that the combination of these techniques provides more accurate results and makes textual information more comprehensible. These approaches are complementary to each other. Statistical approach explores the sentiment and the associations of financial text, while dictionaries provide quantitative data about the volume and the trends of disclosures (Matthies & Coners, 2015).

2.2 Machine learning methods

Guo et al. (2016) suggest machine learning methods such as Neural Network, Support Vector Machine, Naïve Bayes, and semantic analysis. Support Vector Machine is another machine learning method that does not introduce noise in the learning process and that way overfitting problems are avoided. Guo et al. (2016) use a sample of news articles with positive and negative information about companies and they find that Support Vector Machine outperforms Naïve Bayes in terms of out-of-sample forecast accuracy.

In addition, Li (2010) uses Naïve Bayes to quantify narratives from financial reports along with earnings and liquidity. To evaluate the results of Naïve Bayes, he also uses the dictionary methodology.

Furthermore, Neural Network is a highly accurate method that can process vast amount of data. Thomson Reuters News Analytics uses Neural Network in text analytics. Moreover, data from Thomson Reuters News Analytics are extensively used by scholars to investigate the impact of media news on company’s performance and predictability of returns. When Guo et al. (2016) implement Neural Network on their sample they find that both in-sample and out-of-sample predictions provide accurate results.

In addition, a method for semantic analysis is Latent Dirichlet Allocation (LDA) that is very useful for examining the topic distribution and the conceptual content in a business document. Guo et al. (2016) suggest this method for textual analysis of analyst and financial reports. Dyer et al. (2017) use Latent Dirichlet Allocation (LDA) to study certain topics in 10-K filings.

2.3 Document similarity and linear regression models

In case that there is an attempt to identify how similar are the texts, cosine similarity can be implemented.

The definition of cosine similarity is provided by the following equation:

$$\text{Cosine similarity } (d_1, d_2) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2} \sqrt{\sum_i y_i^2}} \quad (1)$$

d_1, d_2 are the documents that are divided into the vectors x and y that include word counts. The values of cosine similarity are from 0 to 1. However, it is a measure that is not extensively used in the fields of finance and accounting (Loughran & McDonald, 2016).

In addition, a widely used method is the implementation of linear regression in time series data to examine the association between textual sentiment of market and performance of firms in stock markets.

This relationship is described by the following equation:

$$y_t = \alpha_0 + \sum_{j=1}^J \alpha_{1,j} y_{t-j} + \sum_{j=0}^J \sum_{f=1}^F \alpha_{2,j}^f X_{t-j}^f + \sum_{j=0}^J \sum_{\alpha=1}^A \alpha_{3,j}^\alpha S_{t-j}^\alpha + \varepsilon_t \quad (2)$$

y is the dependent variable that is usually a variable that reflects the corporate performance on the firm-level or the market-level.

This equation indicates that the dependent variable is formed according to “a α_0 constant term, to $j = 1 \dots J$ lagged dependent terms, $\sum_{j=1}^J \alpha_{1,j} y_{t-j}$, to $j = 0 \dots J$ contemporaneous or lagged terms of $f = 1 \dots F$ control variables in the X vector, to $\sum_{j=0}^J \sum_{f=1}^F \alpha_{2,j}^f X_{t-j}^f$, on $j = 0 \dots J$ contemporaneous or lagged terms of $a = 1 \dots A$ terms in the S vector, to $\sum_{j=0}^J \sum_{\alpha=1}^A \alpha_{3,j}^\alpha S_{t-j}^\alpha$, and to ε_t which is the residual unexplained component.”

However, previous literature indicates that we can examine this regression using as dependent variable the expressed sentiment and as independent variables measures of corporate performance at firm-level and market-level.

Furthermore, in their research Kearney and Liu (2014) refer to vector autoregression (VAR) model that can estimate the interdependencies between sentiment, control variables, and variables of performance on the firm-level and the market-level.

2.4 Methods to measure readability

In the fields of finance and accounting, readability of corporate reports measures the ability of reader to understand the information content of reports. In specific, Loughran and McDonald (2014a) define readability as the capacity of analysts and investors to interpret value-relevant information from financial documents.

The size of the report (natural logarithm of the size of the annual report in megabytes) and number of words in the report (natural logarithm of the number of words of the report) are used to measure readability of financial reports according to Loughran and McDonald (2016).

Moreover, prior research has used several indices to evaluate if financial and accounting disclosures are comprehensible. A widely used index is the Fog Index which consists of two parameters: “complex words” (words that have more than two syllables) in the text and the length of the sentence on average. The Flesch-Kincaid Index estimates the average number of syllables per word and is also used extensively by researchers (Loughran & McDonald, 2016).

Bonsall et al. (2017) suggest a new measure of financial reporting readability, Bog Index that is developed according to plain English attributes that are suggested by SEC. The equation of the Bog Index is the following:

$$\text{Bog Index} = \text{Sentence Bog} + \text{Word Bog} - \text{Pep} \quad (3)$$

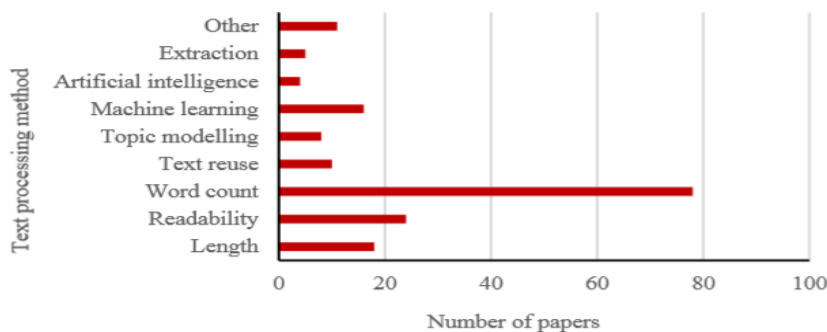
$$\text{Word Bog} = \frac{(\text{Style problems} + \text{Heavy words} + \text{Abbreviations} + \text{Specialist} * 250)}{\text{Number of words}}, \quad (4)$$

$$\text{Sentence Bog} = \frac{(\text{Average Sentence length})^2}{\text{Long Sentence Limit}} \quad (5)$$

$$\text{Finally, Pep} = \frac{(\text{Names} + \text{Interest words} + \text{Conversational}) * 25}{\text{Number of Words}} + \text{Sentence Variety} \quad (6)$$

Some of these attributes are hidden verbs, the use of passive voice, and the length of the sentence. Researchers also use complementary plain English attributes such as number of words and jargon. Bonsall et al. (2017) find that plain English attributes are more accurate measures of readability attributes of financial reports as stated by SEC. Finally, El-Haj et al. (2020) suggest that the most widely used computational linguistics method among researchers is the word count.

Figure 1: Distribution of papers by computational linguistics method



Source: El-Haj et al. (2020).

2.5 Programming languages for textual analysis of sentiment

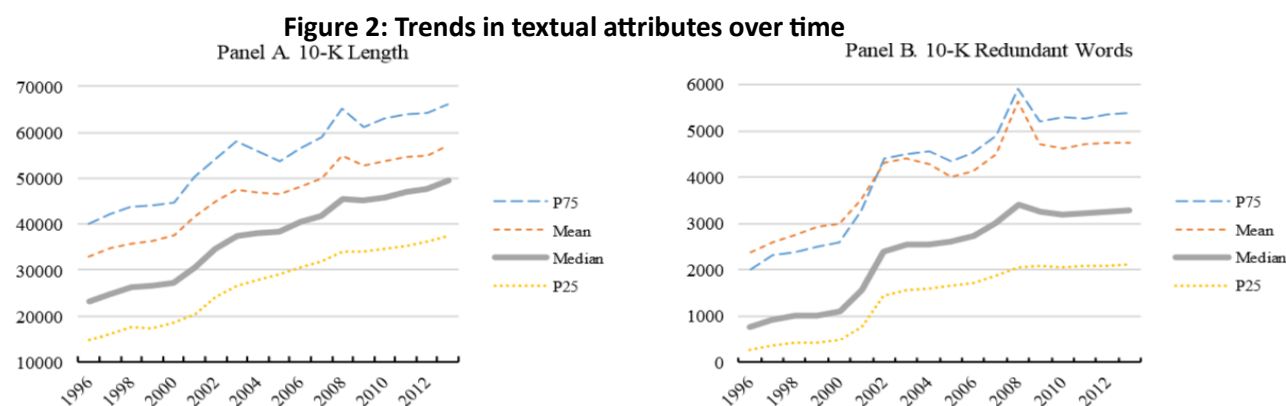
Loughran and McDonald (2016) present various programming languages that are necessary for analyzing financial and accounting disclosures. The choice of programming language is partially subjective and it depends on the preferences of the researcher. However, there are some programming languages that are extensively used and among them is Python.

Statistical programs such as STATA, R, and SPSS can also process financial and accounting narratives. Loughran and McDonald (2016) discourage researchers from creating their own code, as it is possible to produce inaccurate results that are really hard to replicate. On the other hand, they state that researchers must be cautious with programs that are too general, and they do not apply successfully in business context. It is more preferable to use a software that is widely used by researchers of the field.

3 Attributes of textual information in finance and accounting

In this section, we present the trends of characteristics of textual information of financial and accounting disclosures over the years according to established literature in the field.

Dyer et al. (2017) find that the increase in length of the disclosures is due to requirements of SEC and FASB by using Latent Dirichlet Allocation (LDA) to study certain topics in 10-K filings. There are three predominant topics that cause this increase: risk factor, internal controls, and fair value disclosures. Taking into account that the rest of the topics have remained stable over the examined period, the increase of length is probably caused by these three specific topics that are required to be presented thoroughly in 10-K filings. These topics also have a serious impact on trends of textual features. They examine the period 1996-2013 and they conclude that length, boilerplate language, and redundancy have increased while readability and hard information have been reduced.



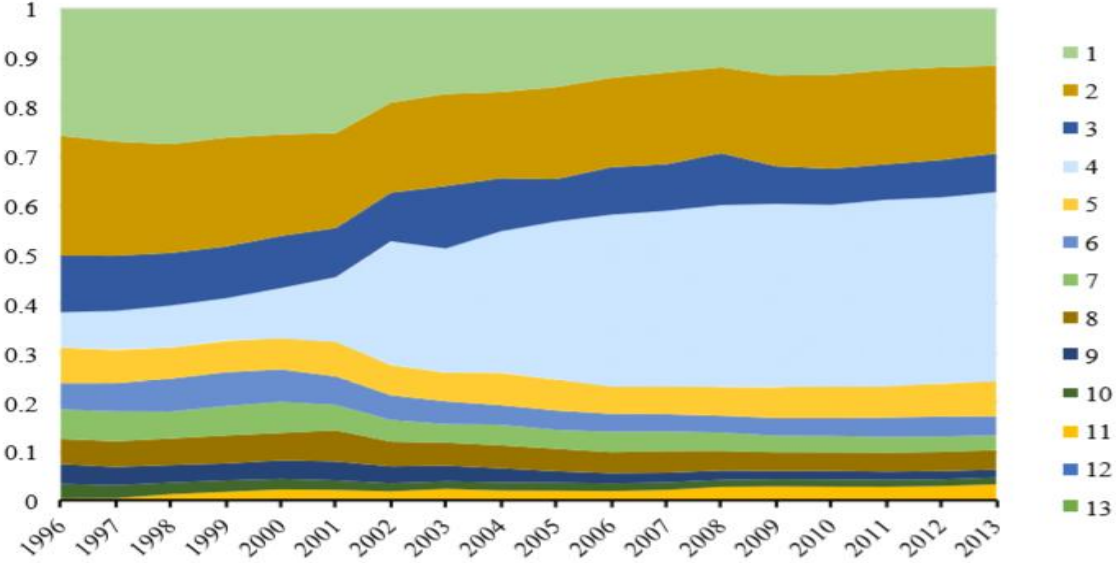


Source: Dyer et al. (2017).

Figure 2 presents the evolution of textual attributes such as length of disclosure, boilerplate, hard information, and readability. The figure clearly illustrates an increase in complexity (Fog index), length, redundancy, boilerplate, and sticky words while there is a decrease in hard information that is contained in 10-Ks and specificity. The next figure presents the median disclosure of various topics as they are formed over the years. It is obvious that compliance with SEC and accounting standards is the topic that is more revealed in annual reports followed by industry-specific disclosure and performance, revenues, and customers.

Figure 3: Median Disclosure Loading by Topic Category

Panel B. Median Disclosure Loading By Topic Category

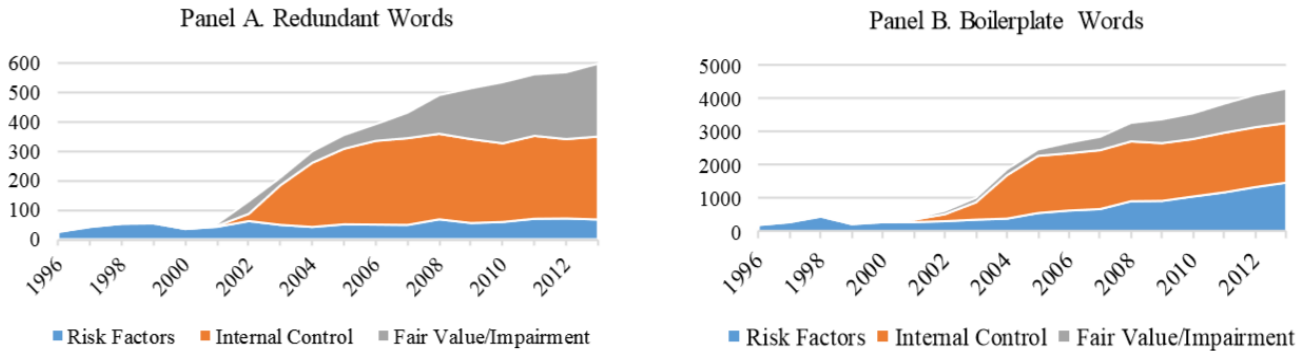


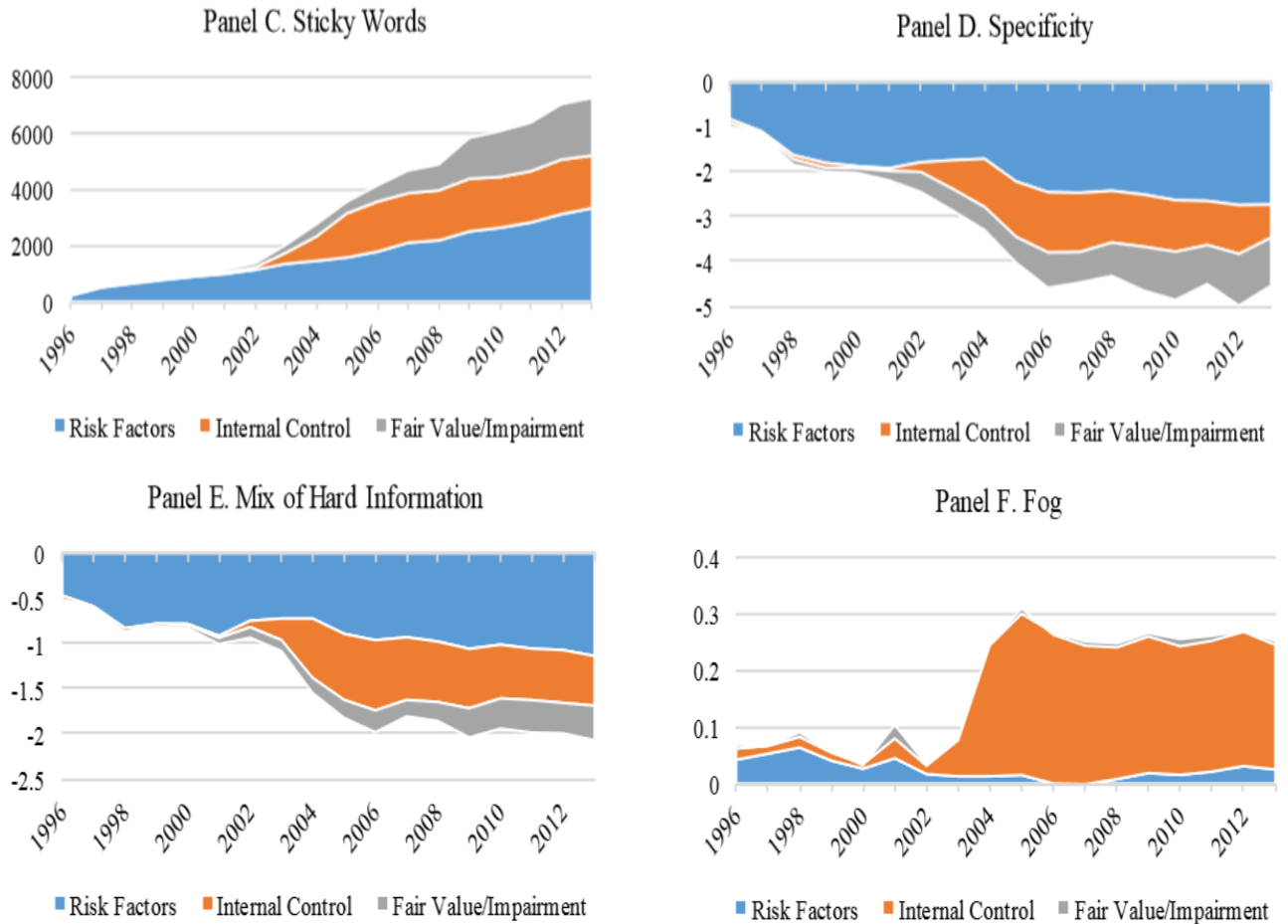
- 1. Performance, Revenues, and Customers
- 2. Industry-Specific Disclosure
- 3. Employees & Executives
- 4. Compliance with SEC & Accounting Standards
- 5. Loans, Debt, Banking
- 6. Business Operations & Strategy
- 7. Stock and Options
- 8. Business Structure & M&A
- 9. Contracts & Legal
- 10. Geographic Location
- 11. Investments, Securities, Derivatives
- 12. Intellectual Property & R&D
- 13. Property and Leasing

Source: Dyer et al. (2017).

In figure 4 Dyer et al. (2017) indicate how the three predominant topics (risk factors, internal control, and fair value/impairment disclosures) affect a variety of textual characteristics over time. They find that textual complexity, redundancy, and boilerplate are highly affected by disclosures of internal controls and specificity, sticky words, and hard information from disclosures of risk factors.

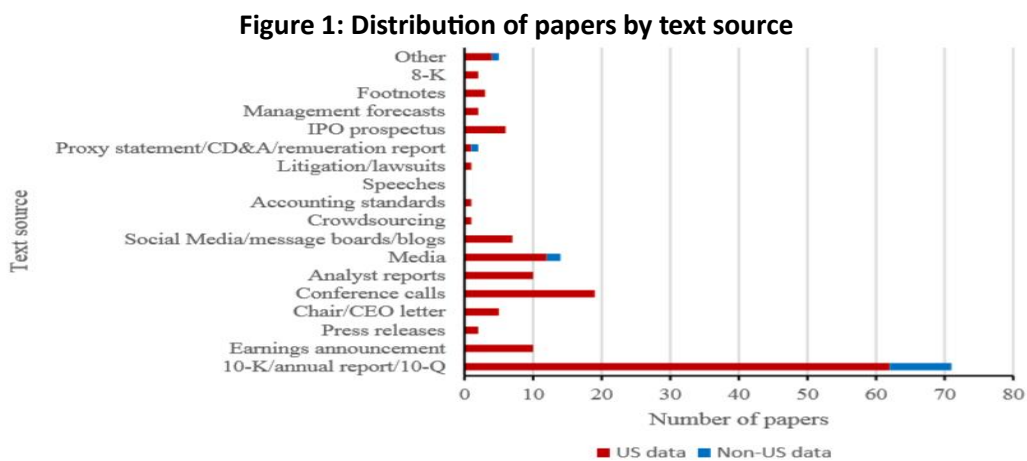
Figure 4: Contribution of the Top 3 Topics to Textual Attributes Over Time



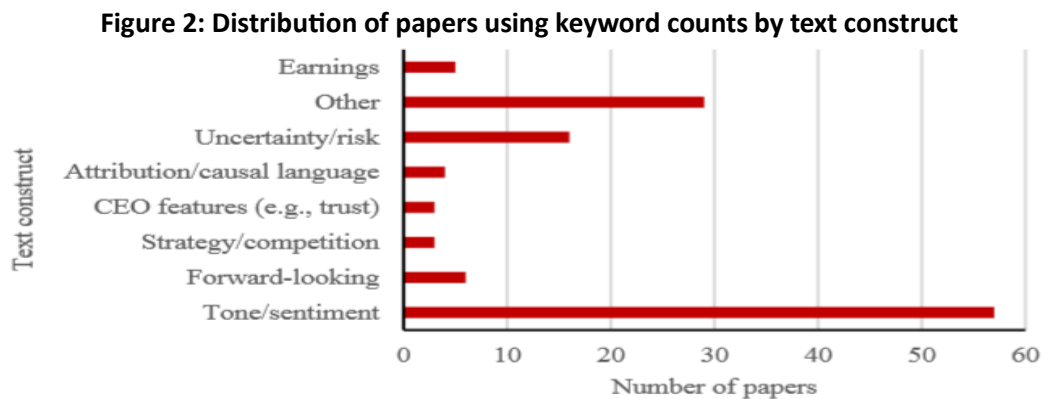


Source: Dyer et al. (2017).

Furthermore, concerning the information sources, figure 5 indicates that the majority of papers in finance and accounting extract data from 10-Ks, annual reports, and 10-Qs. Less popular sources are earnings announcements, conference calls, and media. Finally, figure 6 suggests that most papers use tone/sentiment when analyzing text.



Source: El-Haj et al. (2020).



Source: El-Haj et al. (2020).

4 Textual analysis of sentiment in finance

The last decades, researchers in finance apply extensively textual analysis methods to capture the sentiment of the reports. In specific, Hájek (2018) examines the predictability of annual reports of U.S. firms on abnormal stock returns by using both textual analysis of sentiment and bag-of-words techniques. To implement textual analysis he uses two widely accepted dictionaries: Diction 7.0 and Loughran and McDonald dictionary (Loughran & McDonald, 2011). Textual characteristics are associated with financial indices and neural network to make accurate predictions for abnormal stock returns. In this paper, Gunning Fog Index is also used to examine the readability of annual reports. Latent Semantic Analysis (LSA) alone and Latent Semantic Analysis combined with cosine similarity are used by the researcher to reduce the dimensions of the documents. He finds that the combination of textual analysis of sentiment and machine learning approach provides accurate predictions about market reaction on annual reports.

Balakrishnan et al. (2010) also examine narrative disclosures in 10-K and 10-K405 filings to find if they actually have a predictive ability on stock returns. They find that narrative disclosures and performance indices can provide adequate information for performance prediction. When a firm creates a portfolio based on predictions it achieves important positive returns. Text classification method used by Balakrishnan et al. (2010) can detect information that is not revealed by document - level characteristics of tone, clarity, and risk sentiment such as Fog Index, optimistic-pessimistic tone, and risk sentiment examined in prior literature. Narrative disclosures do not provide additional information to variables that are traditionally used for predictions such as market-to-book ratio, momentum, and size of the firm. However,

they have an impact on investors and in specific, in the way they use price momentum to predict excess returns. The researchers use a text classification method that is based on word extraction from the text, and they develop an algorithmic approach to predict firms' returns. Then, these words are used in predictive models as attributes. A very interesting finding of this paper is that firms with similar quantitative measures such as book-to-market ratio and size have differences in their textual disclosures.

Lee et al. (2014) examine textual disclosures of 8-K documents and they find that information included in 8-K filings has a predictive ability on firm's stock price changes. Textual information increases accuracy by more than 10%. The impact is stronger for the short-term period. Thus, predictions that are based on textual disclosures are more accurate the next day of the financial event. However, the persistence of the impact lasts five days after the announcement of the financial event (Lee et al., 2014).

Furthermore, they find that a model that incorporates textual information performs much better and provides more accurate predictions about stock price changes than a model that includes only financial information. 8-K documents are released when firms face an important financial event such as a bankruptcy, so they provide an insight of various events that have a strong impact on firm's performance. Lee et al. (2014) suggest the construction of a forecasting model that will combine financial and textual information from 8-K documents, and it will provide accurate results for stock prices' movements around a financial event.

Chen et al. (2018) extract sentiment from NASDAQ news articles and they use machine learning techniques to examine the predictability of trading-hour sentiment in equity markets and single-stock option markets. The examined features of single-stocks options are "implied volatility skew, out-of-the-money prices, and implied volatility." They find that contemporaneous sentiment has an impact on single-stock options. In addition, textual sentiment provides incremental information about future stock returns. They state that overnight news are more informative than trading-time news. Overnight news include information that is collected the previous night while trading-time news contain information of the last trading time. Sentiment disagreement results to a strong positive risk premium (Chen et al., 2018).

Feldman et al. (2010) examine the information content of MD&A section in 10-Q and 10-K filings. They classify words into positive and negative categories to examine manager's tone change in comparison with tone in prior reports. They find that tone change affects both short-window excess stock returns, as well as excess stock returns for longer periods. Their findings also indicate association of manager's tone change and instant reaction of market.

Furthermore, they associate optimism and pessimism of MD&A section with predictability of future stock returns. MD&A section provides information beyond accruals and earnings surprises. Moreover, MD&A narratives are more informative for firms that do not have strong information environment, because of their small size and the low number of analysts following them. Finally, tone change has not an important impact on firms with positive earnings surprises and negative accruals.

Moreover, Buehlmaier and Whited (2018) imply textual analysis in annual financial reports of firms to examine the impact of financial constraints on stock returns. The methodology of textual analysis detects financial constraints when there is a clear reference in financial reports. They indicate that Naïve Bayes method is the most appropriate for analyzing the text of annual reports. The content of MD&A section is examined to detect financial constraints. They use access to external financial markets, debt markets, and equity markets to estimate financial constraints and afterwards they examine their impact on stock market. They find that financial constraints are positively associated with stock returns. They indicate that investors demand higher returns for firms that are financially constrained. The results tend to be stronger for debt constraints. Liquid stocks of larger firms tend to return higher earnings to investors, as they have higher risk premium, when their trading strategies are based on financial constraints.

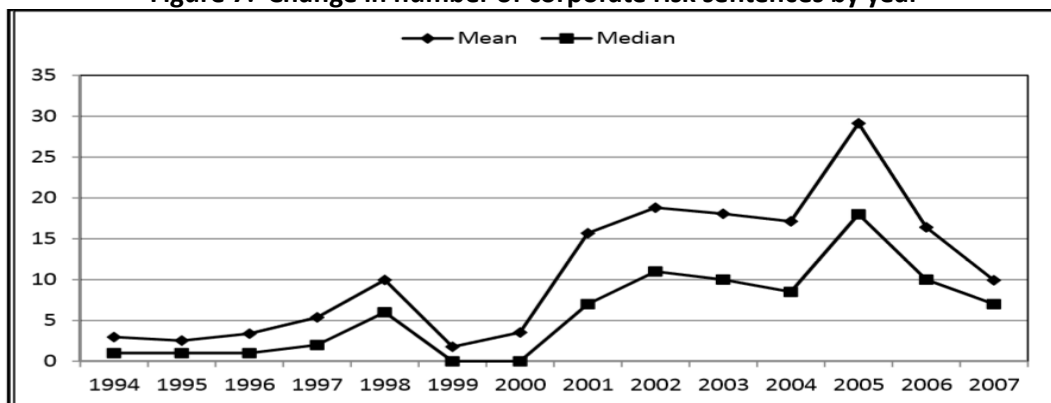
Accounting information contained in financial news is strongly associated with market reaction that is reflected in stock prices (Boudoukh et al., 2013). In this research work, news are examined by tone and by type. Researchers use both a dictionary approach to extract the sentiment of the documents and an analysis of phrase-level trends to extract the tone of the documents. They also use machine learning techniques to automatically classify articles into topics. Stock price volatility is examined on days with no-news, with unidentified news, and identified news. According to the results, days with identified news have an important impact on stock prices' volatility. In these days, there is also high possibility of extreme returns. Finally, tone of daily news is positively correlated with daily stock returns (Boudoukh et al., 2013).

Furthermore, Kravet and Muslu (2013) examine the impact of changes in textual risk disclosures of companies in their 10-K documents on stock market reactions and analyst behavior. There is a positive relationship between annual increases in risk disclosures and increases in trading volume and stock return volatility around corporate filings. Analysts issue more dispersed earnings forecasts and forecast revisions after 10-Ks when risk disclosures are increased. They find that the perceptions of investors about company's risk are associated

positively with textual risk disclosures. Risk perceptions are estimated by trading volume, stock return volatility, and analysts' forecasts. Risk disclosures that provide a huge amount of negative information have not very strong results because it is deemed that managers reveal a lot of negative information to escape the litigation risk. Additionally, disclosures in a firm-level that are different from other firms in the same industry have not so strong results as well. In other words, idiosyncratic firm-specific disclosures are not so useful for report users.

In textual risk disclosures reports' users have information about the range of future operations rather than the level (Kravet & Muslu, 2013). In their paper, they count the number of risk sentences in the document to measure risk disclosures. They examine the impact of changes in risk disclosures on analyst behavior and investors decisions two months before and after the filings.

Figure 7: Change in number of corporate risk sentences by year



Source: Kravet and Muslu (2013).

Risk disclosures also affect changes in volatility of negative daily returns and volume of filings. The results are economically significant relative to the total information presented in firm's documents. Thus, risk disclosures contain incremental information (Kravet & Muslu, 2013).

Figure 7 presents the mean and median change in sentences that contain risk information annually. Information is extracted by 10-K filings of firms from years 1994 to 2007.

Additionally, there is a positive association between risk factor information and pre-disclosure measures that estimate risk. Taking into account the various types of risk (systematic, idiosyncratic, financial etc.) this association is different depending on the type of risk. Stock prices reflect information that is revealed in the risk factor section. The disclosure of unexpected risk has a positive relation with stock market reaction, measured by beta and stock return volatility (Cambell et al., 2014). The unexpected risk factor has also a negative

impact on information asymmetry that is measured by bid-ask-spread. This indicates that information about risk factors of the firms reduces information asymmetry between managers and shareholders. It has also a negative impact on abnormal returns around the release of 10-Ks. In addition, they indicate that risk factor disclosures are very informative, and investors use this information when evaluating firm's value.

Furthermore, 10-K reports of firms have to include a "risk factor" section since 2005. This section must describe to investors and shareholders factors that increase the risk that company faces. Cambell et al. (2014) examine the usefulness of this 10-K section. They find that firms that are more risky include a greater number of risk factors in this section. Additionally, the portion of details that are provided about risks depend on the type of risk that the firm is facing. A very important finding of their research is that information asymmetry, value of firm, systematic risk, and idiosyncratic risk are affected by the information included in the "risk factor" section. The results of this research work hold even after controlling for various parameters that have an impact on the above-mentioned variables such as size, growth, market risk, and changes in analysts' opinions about future cash flows of the firm. Latent Dirichlet Allocation is used in the software that is developed especially for this study to find words that are constantly present in risk factor sections of firms. Moreover, they use three different categories of risk to classify keywords: tax risk, financial risk, and litigation risk.

Kothari et al. (2009), use content analysis to examine the impact of disclosures by managers, analysts, and press on three risk measures: volatility of stock returns, cost of capital, and analyst forecast dispersion.

They find that positive information provided by managers, analysts, and business press reduces these risk factors. Negative information has the opposite effect. Finally, they analyze the effect by source, and they find that the most credible disclosures are those of press as they have the greatest impact on risk variables in comparison with management and analysts' disclosures.

Risk of stock price crash tend to be higher for firms that their 10-K reports are more complex and harder to comprehend (Kim et al., 2019). Annual reports that are less readable are possible to include negative information that management tries to hide from investors. This hidden negative information leads to stock price crash that is accompanied with extreme losses. 10-Ks complexity has a greater impact on stock price risk when information about earnings is consistently negative or temporarily positive, when managers have greater equity option incentives, and when the risk of lawsuits is lower. Risk of stock price crash is positively

associated with manipulation of accruals, in case that the reports are less readable. Kim et al. (2019) use the modified Fog Index to estimate the complexity of language in annual reports. Particularly, they transform the equation of Fog Index, by re-defining complex words:

$$\text{Fog} = (\text{words per sentence} + \text{percentage of complex words}) \times 0.4 \quad (7)$$

Traditionally, complex words are deemed to be these words that have more than three syllables. However, Kim et al. (2019) omit multisyllabic words that are understandable in business context.

Moreover, larger 10-K filings including greater amount of uncertain and weak modal words lead to a greater possibility of stock price crash in the future (Ertugrul et al., 2017). Researchers find a positive association between managerial information hoarding and readability and tone ambiguity of public corporate documents. Managers that issue less readable financial reports are perceived to hide negative information. Furthermore, unfavorable disclosures are gathered gradually and after a tipping point, they are revealed. Thus, firms with less readable reports have higher stock price crash risk. Shareholders of firms with more complex and fuzzy annual reports are burdened with higher cost of financing from external sources, apart from less transparent disclosures. Firms with ambiguous tone in their financial reports are considered that contain high-information risk and that is priced by creditors (Ertugrul et al., 2017). Researchers use as a readability measure the size of 10-K filings.

Li and Zhao (2015) also examine the complexity of annual reports by using as a complexity measure financial reports' size. Their survey sheds light to the association between complexity of 10-Ks and uncertainty that is reflected in volatility. They find that volatility dynamics is hump shaped. The first two to four weeks after the issue of financial reports volatility increases by a small percentage. The following six weeks is reduced significantly. Corporates with larger annual reports tend to incur this phenomenon in a greater extent. This indicates that information complexity results to higher uncertainty in the short-term but in the long-term investors comprehend the content of complex financial reports and uncertainty vanishes. By using bid-ask spread and Amihud's illiquidity measure, the researchers estimate information asymmetry and state that firms with more complex financial disclosures experience larger decreases in information asymmetry in the long horizon (Li & Zhao, 2015).

Bushee et al. (2018) examine the two components of linguistic complexity: obfuscation and information that both have diverse effects on information asymmetry. They examine earnings conference calls of companies by quarter and the information asymmetry that

follows the call.

There is a negative relationship between information factor of linguistic complexity and information asymmetry and a positive relationship between obfuscation factor and information asymmetry. Thus, they suggest that linguistic complexity must be divided in two components (information and obfuscation) to better examine its impact on information asymmetry.

This method of the examination of two different components of linguistic complexity of earnings conference calls provides greater details about firms that report losses, and therefore, not all loss firms disclose complex earnings conference calls with the intention to obfuscate information.

Finally, Huang, H.A. et al. (2014) find that textual disclosures of analyst reports provide investors with incremental information beyond information included in earnings forecasts, target prices, and stock recommendations. Text of analyst reports is a useful tool for investors as it helps them digest the information included in sources mentioned above. Analysts' reports have a stronger impact on investors' decisions when they include unfavorable disclosures. Investors' reaction to negative quantitative measures is stronger when the text of analysts' reports is more negative.

The same holds for positive quantitative measures and more positive textual disclosures of analysts. Additionally, analysts' reports are more important when they provide details for non-financial topics and when the structure is more accurate and concise (Huang, H.A. et al., 2014). Analysts' reports have predictive ability and can shed light to earnings growth for a period of five years after the release of the report. The researchers use a Naïve Bayes approach to analyze the large corpus of analysts' reports.

5 Textual analysis of sentiment in accounting

In this section, we present numerous empirical studies which utilize textual and financial analysis on accounting disclosures. The analysis of information content of accounting disclosures is an emerging field in the accounting research and therefore, many studies in the recent years attempt to examine it.

Previous research in the field indicates that the Management Discussion and Analysis section included in 10-K and 10-Q filings of firms can be very informative when we analyze the forward-looking statements (FLS) (Li, 2010). In his paper, Li uses a Naïve Bayes algorithm to process qualitative information. By using this statistical approach of content analysis, he classifies the tone and the content of FLS. Particularly, he classifies sentences in four different

tone categories: negative, positive, uncertain, and neutral. TONE equals 1, if the sentence is predicted to be positive, 0 if it is predicted to be neutral, and -1 if it is predicted to be uncertain or negative. Naïve Bayes predicts the average tone of all the FLS in a 10-K or a 10-Q filing for each company i in a quarter j and that is the MD&A tone of its filing.

He concludes that companies with “better current performance, lower accruals, smaller size, lower market-to-book ratio, less return volatility, lower MD&A Fog index, and longer history” have forward-looking statements with positive content. Furthermore, this research work sheds light to the association between the content of FLS and future earnings and liquidity of the company. FLS tone can provide additional information about company’s future performance. Specifically, he finds that the average tone of FLS is positively correlated with future earnings of the company, even when control variables of future performance are used. Moreover, the tone of Management Discussion and Analysis section plays a critical role in accrual anomaly. When the content of Management Discussion and Analysis section is informative about the future impact of accruals on firms’ earnings, future returns are not affected by accruals, accruals mispricing is reduced, and market operates more efficiently. In addition, he uses dictionaries as a complementary method of content analysis, and he suggests that they cannot successfully predict future earnings of firms. He uses Diction, General Inquirer, Linguistic Inquiry, and the Word Count and he indicates that these dictionaries are not suitable for analyzing business documents.

Davis and Tama-Sweet (2012) examine textual content of earnings press releases and they compare it with the content of Management Discussion & Analysis Section. They study the differences of managers’ use of language when they describe firm performance in earnings press releases and Management Discussion & Analysis Section. Reaction of market is stronger for textual disclosures of earnings press releases in comparison with MD&A in annual reports. In earnings press releases there are less pessimistic statements than in MD&A section of SEC filings. They develop a measure of total ratio of pessimistic words in earnings press releases relative to pessimistic words in MD&A section. This measure has a negative association with “the intensity of managers’ strategic reporting incentives” (Davis & Tama-Sweet, 2012). There is also a negative association between pessimistic words in MD&A and future performance of the company, after controlling for pessimistic words in these disclosures. Thus, pessimistic tone in MD&A section provides incremental information to earnings press release.

Earnings press releases are prior to MD&A and investors, analysts, and financial press are strongly interested in them. Because of that, managers tend to disclose more optimistic information in earnings press releases as they have greater impact on market participants.

Moreover, earnings press releases contain less pessimistic language in case of high growth firms and firms that exactly meet or just beat earnings benchmarks in the current quarter. The same holds for firms that persistently meet or just beat analysts' earnings forecasts. Unfavorable disclosures in MD&A section have a predictive ability on firm's future performance (Davis & Tama-Sweet, 2012).

Allee and DeAngelis (2015) examine the role of tone dispersion to shed light to qualitative disclosures and their association to voluntary disclosures of managers and the reaction of information users to these disclosures. Tone dispersion shows the degree to which tone words are spread evenly in the financial text. Tone dispersion, both positive and negative, is correlated with current and future performance of the firm. Tone dispersion is also correlated with decisions of managers about financial reporting and their actions and incentives to create a favorable image for the firm (Allee & DeAngelis, 2015). They investigate market and analyst reaction to the textual content of conference calls, and they find that manager's narratives have an impact on information that is received by investors and analysts. Information users react negatively to more dispersed negative tone, while the opposite holds for positive tone. Authors use the financial word list developed by Loughran and McDonald (2011) to capture the positive and negative tone dispersion of conference calls.

Huang, X. et al. (2014) also investigate the tone of managers. However, they examine earnings press releases and not annual reports because of their higher frequency. They test the hypothesis that firms manage the tone of their disclosures, and they study the impact of tone management on investors. Abnormal positive tone is higher for firms that exceed past earnings or analysts' forecasts', for profitable firms, and for firms that are going to be merged or acquired or they are going to issue a new equity. ABTONE hides unfavorable information about fundamentals of a firm, and it is positively correlated with low future earnings and cash flows in a one to three-years horizon. Manipulation of tone in earnings press releases eventually provide unreliable information to investors, as managers tend to hide adverse information about poor future performance of the firm.

Tetlock et al. (2008) indicate that qualitative information of articles can predict future performance of corporates on firm-level. They use the fraction of negative words, and they find that negative corporate information in financial news is a signal of negative future earnings. Negative information about firms' fundamentals in news can provide more accurate estimations about future earnings in comparison with other types of financial information.

Furthermore, Lang and Stice-Lawrence (2015) use a large sample of 15,000 non-U.S. companies from 42 different countries and they examine the complexity, the length of

disclosures of these firms, the existence of boilerplate, and the extent that the reports of non-U.S. firms are comparable with those of U.S. firms. Comparability is estimated by using the cosine similarity method and complexity is examined by using the Fog Index. Textual features of annual reports are associated with the number of analysts following the firm, liquidity, and institutional ownership (Lang & Stice-Lawrence, 2015). Specifically, companies with greater liquidity, higher institutional ownership, and more analysts that follow them have more comparable and larger reports. These firms' reports contain less repeatability. They also indicate that the obligatory establishment of IFRS enhanced the quality of financial reporting in sense that boilerplate declined and the length of disclosures and comparability among reports of U.S. and non-U.S. firms increased. Companies that benefited after IFRS adoption in terms of reporting quality achieved greater increases in their liquidity, analyst following, and institutional ownership. Lang and Stice-Lawrence (2015) shed light to the change of topics discussed in financial reports after mandatory IFRS adoption. They also implement the natural language processing method of Latent Dirichlet Allocation (LDA), and they find that IFRS adoption caused the increase of information about "financial instruments and derivatives", "executive compensation", and "employee stock options." Quality of disclosures increases when firms' auditing is stronger, for example when a firm is audited by a Big 5 audit firm.

In addition, empirical evidence states that Financial Reporting Complexity (FRC) has a positive association with accounting expertise of board of directors and auditors in firms (Chychyla et al., 2019). The unfavorable results of reporting that are produced by FRC are reduced with the contribution of accounting experts. The researchers consider as unfavorable results that are mitigated because of accounting expertise: the "internal control weakness" and "SEC comment letters." They measure Financial Reporting Complexity by examining accounting standards. In specific, they develop a measure of complexity that is based on the length of accounting standards that are associated with each item of 10-K filings. This approach estimates the effort that is put on the preparation of annual reports based on the length of accounting standards, rather than the effort of the readers to understand them.

Moreover, firms with financial constraints adopt tax planning strategies that are rather aggressive (Law & Mills, 2015). Firms with more negative words, are financially constrained and they have higher unrecognized tax benefit balances, thus they follow more aggressive tax strategies. A measure for financial constraint is the amount of negative words in financial reports. In their research, Law and Mills (2015), indicate that financially constrained firms imply more aggressive tax strategies, and this can be detected by a) current and future effective tax rates that are lower both in the short-term and in the long-term, b) higher current

and future tax benefits that have not been recognized, c) Internal Revenue Service suggesting greater amount of audit adjustments and, finally d) material operations are associated with the use of tax havens. The researchers also examine the tendency of firms to avoid their tax responsibilities as a result of an exogenous shock. That is the closure of local banks which is an external source that has an impact on financial constraints of firms. Qualitative disclosures can provide incremental information beyond the analysis of accounting numbers and are a very useful approach to examine and predict tax-aggressive behavior of firms (Law & Mills, 2015).

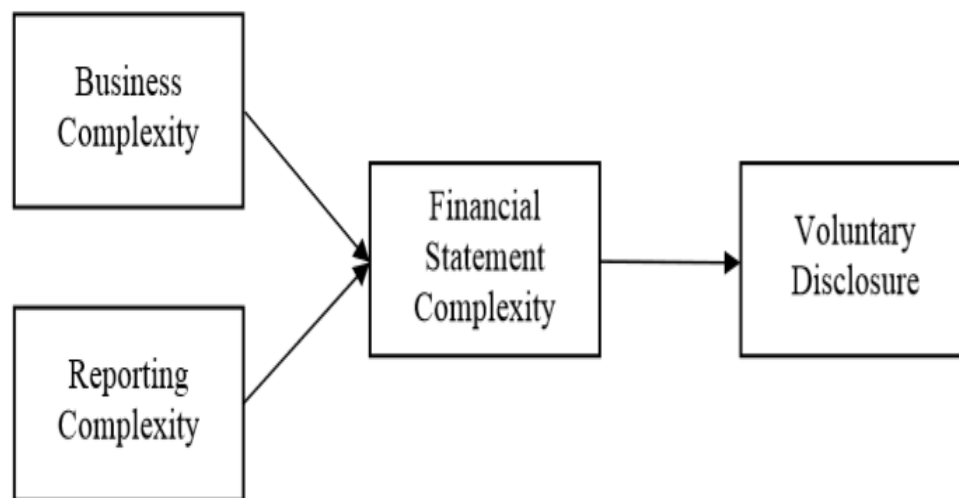
Bodnaruk et al. (2015) also examined financial constraint firms by parsing text of annual reports. Specifically, they create a lexicon based on words related to constraint. They find that widely used accounting measures of constraint have a little association with this lexicon. This lexicon, that detects the frequency of constraining words in a 10-K text, performs better than accounting measures and it has greater predictability on firm's liquidity events. "Dividend omissions or increases, equity recycling, underfunded pensions" are better predicted by this lexicon approach. Bodnaruk et al. (2015) include 184 constraining words in their list, and they analyze the entire 10-K file to detect constraining words.

In addition, Smugereski (2019) examines the association between linguistic style that managers use in their disclosures and firm's policies. Specifically, he examines earnings conference calls and the sentiment expressed by managers. Temperament of managers has an impact on firm's policies. Higher debt to equity ratios are observed in firms with more aggressive managers. Moreover, managers with more aggressive linguistic style prefer financing that is deemed riskier. He uses a Python program to extract sentiment and create word counts. In this survey, sentiment word counts in the master dictionary file are based on Loughran and McDonald dictionary (Loughran & McDonald, 2011). Harvard General Inquirer word list is also used because of its application in the field of psychology. Smugereski (2019) wants to capture the personality of CEO's because of that he uses this word list. In specific, he focuses on hostile word list to find if the personality of the CEO is actually aggressive.

Complex financial information has a negative impact on information environment of the firm (Guay et al., 2016). Managers may use voluntary disclosure to decrease the negative effect of complex information. There is a positive association between financial statement complexity and voluntary disclosure. A stronger impact is observed when companies have outside monitoring and when liquidity decreases around the filing of financial statements. External monitoring is measured by the number of analysts that follow the firm. However, the relationship is not that strong for companies that underperform and have a higher level of

earnings management (Guay et al., 2016). Firms that adopted more complex disclosures because of the accounting standards disclose voluntarily more information. Furthermore, complexity that arises from the intentional choice of managers to hide adverse information is expected to have no relation with voluntary disclosure. Complex information that arises from the adoption of accounting standards (such as SFAS 133 and SFAS 157) or multiple segments of business is expected to be associated positively with voluntary disclosure.

Figure 8: The relation between financial statement complexity and voluntary disclosure



Source: Guay et al. (2016).

Guay et al. (2016) estimate complexity of financial statements by using readability and length of 10-Ks. Voluntary disclosures are estimated by the frequency of manager's forecasts one year after the filing of the 10-K. They find that the complexity of a 10-K has a greater association with voluntary disclosure after 10-K filing than before the 10-K filing.

6 Recent research and future opportunities

Recently, researchers have employed artificial intelligence techniques to detect the sentiment of the text (Kusal et al., 2021). The evolution of artificial intelligence has a significant impact on the preparation of financial and accounting disclosures. In specific, AI affects the decision of managers to make disclosures that are easier to be processed and to use tone and content that is positively interpreted by algorithms. Moreover, this pattern is more intense for firms that either may increase their benefits (obtain external financing) or decrease their expenses (reduce legal risk) by managing the sentiment of their disclosures (Cao et al., 2023). In addition, recent research implements artificial intelligence techniques for textual analysis

to interpret the information content of text (Hong et al., 2022). Furthermore, Kusal et al. (2021) state that artificial intelligence is used to detect the sentiment of text by combining computations and linguistics methods.

Teoh (2018) examines new, innovative datasets that can be used in research such as Google Trends data and social network and social media data from Twitter and LinkedIn. These datasets provide many potential possibilities to researchers but also include several challenges. Textual data from corporate documents can be used as auditing tools to detect low quality of financial reporting to prevent earnings management.

New datasets may shed light to investors' response to financial and accounting information and to the decisions of managers to reveal or hide information from capital markets. A very important advantage of the use of new datasets in finance and accounting research and especially textual information is that it can solve econometric problems like causality (Teoh, 2018).

El-Haj et al. (2020) state that finance and accounting research do not use extensively computer linguistic methods in comparison with other fields of research. Furthermore, sentiment extraction from words is partially examined in finance and accounting research. Practical relevance is also limited because of issues in structure (El-Haj et al., 2020). Computer linguistics techniques are supplementary tools to financial analysis. They suggest four tools of computational linguistics that should be used extensively in future research: "entity recognition (NER), summarization, semantics, and corpus linguistics" (El-Haj et al., 2020).

7 Conclusion and discussion

In this paper, we have examined the literature in the field of textual analysis of sentiment in finance and accounting. We attempt to shed light on four key areas of the research: methodologies for measuring textual sentiment in finance and accounting, characteristics of textual information of financial and accounting disclosures and their evolution over the years, textual analysis of sentiment in finance research as well as in accounting research.

Regarding the methodologies of textual analysis of financial and accounting information, we find that dictionaries that use word lists to capture the sentiment of the documents is a very widely used method in the field. Loughran and McDonald dictionary (Loughran & McDonald, 2011) is a very popular dictionary and is also created especially for the analysis of sentiment of business documents. Furthermore, prior literature suggests that machine learning methods can be implemented for textual analysis such as Neural Network, Support Vector Machine, Naïve Bayes, and semantic analysis (Guo et al., 2016). Moreover, when a

researcher intends to examine specific topics from financial or accounting disclosures he could use Latent Dirichlet Allocation (LDA) to study certain topics in 10-K filings (Dyer et al., 2017). In addition, cosine similarity is a methodology that can be implemented to reveal similarity of disclosures (Loughran & McDonald, 2016). For measuring the complexity of the financial and accounting disclosures established research indicates various techniques such as: the size of the report, the number of words, Fog Index, Flesch-Kincaid Index, and Bog Index (Loughran & McDonald, 2016; Bonsall et al., 2017). Concerning the programming languages that can be used, researchers focus mainly in Python programming language (Loughran & McDonald, 2016).

Furthermore, we have reviewed the literature focusing on attributes of textual information of financial and accounting disclosures and we have discovered that length, boilerplate language, and redundancy of business documents have increased while readability and hard information have decreased (Dyer et al., 2017). Moreover, annual disclosures focus more on compliance with SEC and accounting standards (Dyer et al., 2017). Regarding the information sources, we find that the majority of researchers use data from 10-Ks, annual reports, and 10-Qs, while textual analysis focus on the tone/sentiment of disclosures (El-Haj et al., 2020).

In addition, literature review of textual analysis in finance reveals that textual sentiment of disclosures has potentially strong impacts on predictions concerning market reaction on annual reports (Hájek, 2018), stock price changes (Lee et al., 2014), single-stock options, and future stock returns (Chen et al., 2018; Feldman et al., 2010). Furthermore, textual information is significantly associated with stock returns (Buehlmaier & Whited, 2018) and stock prices (Boudoukh et al., 2013). Prior research in textual analysis in finance indicates that textual risk disclosures have an impact on trading volume and stock return volatility (Kravet & Muslu, 2013), stock market reaction, abnormal returns, systematic risk, and idiosyncratic risk (Cambell et al., 2014). Moreover, textual information has a statistically significant relation with volatility of stock returns, cost of capital, and analyst forecast dispersion (Kothari et al., 2009). Prior research reveals that the complexity of textual information has a strong correlation with stock price crash (Kim et al., 2019; Ertugrul et al., 2017), volatility (Li & Zhao, 2015), and information asymmetry (Bushee et al., 2018).

Literature review of textual analysis in accounting indicates that textual sentiment is significantly associated with future liquidity (Li, 2010; Lang & Stice-Lawrence, 2015) and earnings/performance (Li, 2010; Davis & Tama-Sweet, 2012; Allee & DeAngelis, 2015; Huang, X. et al., 2014; Tetlock et al., 2008). Furthermore, prior research associates complexity of

accounting information with accounting expertise of board of directors and auditors (Chychyla et al., 2019) and voluntary disclosures of managers (Guay et al., 2016). In addition, textual sentiment is strongly correlated with financial constraints of firms (Law & Mills, 2015; Bodnaruk et al., 2015). Finally, sentiment expressed by managers has a significant impact on firm's financing policies (Smugereski, 2019).

In conclusion, we identify the following areas for future research. First, textual analysis of the sentiment of financial and accounting information could be enhanced by implementing artificial intelligence techniques. The evolution of artificial intelligence provides plenty of opportunities for textual analysis in the field. Second, there is a variety of new, innovative datasets that researchers could explore such as Google Trends data and social network and social media data from Twitter and LinkedIn. Finally, computer linguistics techniques could be utilized more systematically to increase the efficiency of financial analysis.

References

- Allee, D. K., & DeAngelis, D. M. (2015). The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion. *Journal of Accounting Research*, 53(2), 241-274. <https://doi.org/10.1111/1475-679X.12072>
- Al-Shaer, H., Salama, A., & Toms, S. (2017). Audit committees and financial reporting quality: Evidence from UK environmental accounting disclosures. *Journal of Applied Accounting Research*, 18(1), 2-21. <https://doi.org/10.1108/JAAR-10-2014-0114>
- Balakrishnan, R., Qiu, Y. X., & Srinivasan, P. (2010). On the predictive ability of narrative disclosures in annual reports. *European Journal of Operational Research*, 202(3), 789-801. <https://doi.org/10.1016/j.ejor.2009.06.023>
- Bodnaruk, A. Loughran, T., & McDonald, B. (2015). Using 10-K text to gauge financial constraints. *Journal of Financial and Quantitative Analysis*, 50(4), 623-646. <https://doi.org/10.1017/S0022109015000411>.
- Bonsall, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2/3), 329-357. <https://doi.org/10.1016/j.jacceco.2017.03.002>
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2013). *Which news moves stock prices? A textual analysis* (NBER Working Paper No. 18725). National Bureau of Economic Research. <https://doi.org/10.3386/w18725>
- Buehlmaier, M., & Whited, T. (2018). Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, 31(7), 2693-2728. <https://doi.org/10.1093/rfs/hhy007>
- Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1), 85-121. <https://doi.org/10.1111/1475-679X.12179>
- Cambell, L. J., Chen, H., Dhaliwal, S. D., Lu, H., & Steele B. L. (2014). The information content of mandatory risk factor disclosures in corporate filings. *Review of Accounting Studies*, 19(1), 396-455. <https://doi.org/10.1007/s11142-013-9258-3>
- Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. *The Review of Financial Studies*, 36(9), 3603-3642. <https://doi.org/10.1093/rfs/hhad021>

- Chen, C., Fengler, R. M., Härdle, K. W., & Liu, Y. (2018). *Textual sentiment, option characteristics, and stock return predictability* (IRTG 1792 Discussion Paper No. 2018-030). Humboldt-Universität zu Berlin. <https://ssrn.com/abstract=3210585>
- Chychyla, R., Leone, A., & Minutti-Meza, M. (2019). Complexity of financial reporting standards and accounting expertise. *Journal of Accounting and Economics*, 67(1), 226-253. <https://doi.org/10.1016/j.jacceco.2018.09.005>
- Davis, K. A., & Tama-Sweet, I. (2012). Managers' Use of Language across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A. *Contemporary Accounting Research*, 29(3), 804-837. <https://doi.org/10.1111/j.1911-3846.2011.01125.x>
- DeZoort, F. T., Hermanson, D. R., Archambeault, D. S., & Reed, S. A. (2002). Audit Committee Effectiveness: A Synthesis of the Empirical Audit Committee Literature. *Journal of Accounting Literature*, 21, 38-75. <https://digitalcommons.kennesaw.edu/facpubs/1495/>
- Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2/3), 221-245. <https://doi.org/10.1016/j.jacceco.2017.07.002>
- Eachempati, P., Srivastava, P. R., Kumar, A., Tan, K. H., & Gupta, S. (2021). Validating the impact of accounting disclosures on stock market: A deep neural network approach. *Technological Forecasting and Social Change*, 170. <https://doi.org/10.1016/j.techfore.2021.120903>
- El-Haj, M., Alves, P., Rayson, P., Walker, M., & Young, S. (2020). Retrieving, classifying and analyzing narrative commentary in unstructured (glossy) annual reports published as PDF files. *Accounting and Business Research*, 50(1), 6-34. <https://doi.org/10.1080/00014788.2019.1609346>
- Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis*, 52(2), 811-836. <https://doi.org/10.1017/S0022109017000187>
- Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, 15, 915-953. <https://doi.org/10.1007/s11142-009-9111-x>
- Frazier, K. B., Ingram, R. W., & Tennyson, B. M. (1984). A methodology for the analysis of narrative accounting disclosures. *Journal of Accounting Research*, 318-331. <https://doi.org/10.2307/2490713> or <https://www.jstor.org/stable/2490713>
- Guay, W., Samuels, D., & Taylor, D. (2016). Guiding through the Fog: Financial statement complexity and voluntary disclosure. *Journal of Accounting and Economics*, 62(2/3), 234-269. <https://doi.org/10.1016/j.jacceco.2016.09.001>
- Guo, L., Shi, F., & Tu, J. (2016). Textual analysis and machine learning: Crack unstructured data in finance and accounting. *The Journal of Finance and Data Science*, 2(3), 153-170. <https://doi.org/10.1016/j.jfds.2017.02.001>
- Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. *Neural Computing and Applications*, 29(7), 343-358. <https://doi.org/10.1007/s00521-017-3194-2>
- Hong, M.-H., Marsh, L. A., Feuston, J. L., Ruppert, J., Brubaker, J. R., & Szafr, D. A. (2022). *Scholastic: Graphical human-AI collaboration for inductive and interpretive text analysis*. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (pp. 1-12). Association for Computing Machinery. <https://dl.acm.org/doi/abs/10.1145/3526113.3545681>
- Huang, H. A., Zang, Y. A., & Zheng, R. (2014). Evidence on the Information Content of Text in Analyst Reports. *The Accounting Review*, 89(6), 2151-2180. <https://doi.org/10.2308/accr-50833>

- Huang, X., Teoh, S. H., & Zhang, Y. (2014). Tone management. *The Accounting Review*, 89(3), 1083-1113.
https://www.jstor.org/stable/pdf/24468328.pdf?casa_token=TrMf4oQzY1EAAAAA:YpsMAXNOblmHrITvDgXEa_ILG3-Zq8mT_XnHAXMN5tlbJ-AQf6BkPZqCmiSVHBqPKSq6li1E4uKh2UvtmckS-GpMaZZiz1CrnBV1TY7OpgteJyd98M5imQ
- Iker, H. P. (1974). SELECT: A computer program to identify associationally rich words for content analysis. I. Statistical results. *Computers and the Humanities*, 8(5/6), 313-319.
<https://www.jstor.org/stable/30199711>
- Ingram, R. W., & Frazier, K. B. (1980). Environmental performance and corporate disclosure. *Journal of Accounting Research*, 614-622.
https://www.jstor.org/stable/pdf/2490597.pdf?casa_token=JgdHyFr3m5sAAAAA:yYnig5bVlz2Z0AH7hWP28B-UGJxxzMnVR3thbby9Vwvp1d8AiJBDkAEmfeUbEt46OmvT652D_6aW3A_EVlRghQmgPytYkNNCWU-ecm0zO6uSd3lv0zvf5hw
- Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33, 171-185.
<https://doi.org/10.1016/j.irfa.2014.02.006>
- Kim, C., Wang, K., & Zhang, L. (2019). Readability of 10-K reports and stock price crash risk. *Contemporary Accounting Research*, 36(2), 1184-1216.
<https://doi.org/10.1111/1911-3846.12452>
- Kothari, S. P., Li, X., & Short E. J. (2009). The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study Using Content Analysis. *The Accounting Review*, 84(5), 1639-1670.
<https://doi.org/10.2308/accr.2009.84.5.1639>
- Kravet, T., & Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, 18(4), 1088-1122. <https://doi.org/10.1007/s11142-013-9228-9>
- Kusal, S., Patil, S., Kotecha, K., Aluvalu, R., & Varadarajan, V. (2021). AI Based Emotion Detection for Textual Big Data: Techniques and Contribution. *Big Data and Cognitive Computing*, 5(3). <https://doi.org/10.3390/bdcc5030043>
- Laine, M. (2010). Towards sustaining the status quo: Business talk of sustainability in Finnish corporate disclosures 1987-2005. *European Accounting Review*, 19(2), 247-274.
<https://doi.org/10.1080/09638180903136258>
- Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2/3), 110-135.
<https://doi.org/10.1016/j.jacceco.2015.09.002>
- Law, K. K. F., & Mills, L. F. (2015). Taxes and financial constraints: Evidence from linguistic cues. *Journal of Accounting Research*, 53(4), 777-819. <https://doi.org/10.1111/1475-679X.12081>
- Lawrence, A. (2013). Individual investors and financial disclosure. *Journal of Accounting and Economics*, 56(1), 130-147. <https://doi.org/10.1016/j.jacceco.2013.05.001>
- Lee, H., Surdeanu, M., MacCartney, B., & Jurafsky, D. (2014). On the importance of text analysis for stock price prediction. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odiijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1065_Paper.pdf
- Li, F. (2010). The information content of forward-looking statements in corporate filings- A naïve Bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049-1102. <https://doi.org/10.1111/j.1475-679X.2010.00382.x>

- Li, J., & Zhao, X. (2015). *Complexity and information content of financial disclosures: Evidence from evolution of uncertainty following 10-K filings* [Working paper]. <https://ssrn.com/abstract=2516622>
- Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- Loughran, T., & McDonald, B. (2014a). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4), 1643-1671. <https://doi.org/10.1111/jofi.12162>
- Loughran, T., & McDonald, B. (2014b). Regulation and financial disclosure: The impact of plain English. *Journal of Regulatory Economics*, 45, 94-113. <https://doi.org/10.1007/s11149-013-9236-5>
- Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, 54(4), 1187-1230. <https://doi.org/10.1111/1475-679X.12123>
- Matthies, B., & Coners, A. (2015). Computer-Aided Text Analysis of Corporate Disclosures - Demonstration and Evaluation of Two Approaches. *The International Journal of Digital Accounting Research*, 15, 69-98. <https://pdfs.semanticscholar.org/7b53/4d07ef02c1809de3268a2739bd9d975a031a.pdf>
- Muslu, V., Mutlu, S., Radhakrishnan, S., & Tsang, A. (2019). Corporate social responsibility report narratives and analyst forecast accuracy. *Journal of Business Ethics*, 154(4), 1119-1142. <https://doi.org/10.1007/s10551-016-3429-7>
- Nazari, J. A., Hrazdil, K., & Mahmoudian, F. (2017). Assessing social and environmental performance through narrative complexity in CSR reports. *Journal of Contemporary Accounting and Economics*, 13(2), 166-178. <https://doi.org/10.1016/j.jcae.2017.05.002>
- Smugereski, J. M. (2019). *Reading between the lines: CEO temperament measured by textual analysis and firm policy* (Honors thesis, University of New Hampshire). Scholars' Repository. <https://scholars.unh.edu/honors/467>
- Tetlock, C. P., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3), 1437-1467. <https://doi.org/10.1111/j.1540-6261.2008.01362.x>
- Teoh, H. S. (2018). The promise and challenges of new datasets for accounting research. *Accounting, Organizations and Society*, 68-69, 109-117. <https://doi.org/10.1016/j.aos.2018.03.008>