



REVISTA DE CONTABILIDAD

SPANISH ACCOUNTING REVIEW

revistas.um.es/rcsar



Textual analysis and sentiment analysis in accounting

Juan L. Gandía^a, David Huguet^b

a, b) Departamento de Contabilidad, Facultad de Economía, Universidad de Valencia, Valencia, España.

^aCorresponding author.

E-mail address: juan.l.gandia@uv.es

ARTICLE INFO

Article history:
Received 1 July 2019
Accepted 3 February 2020
Available online 1 July 2021

JEL classification:
M41
M42

Keywords:
Textual analysis
Sentiment analysis
Qualitative analysis
Signalling theory

Códigos JEL:
M41
M42

Palabras clave:
Análisis textual
Análisis de sentimientos
Análisis cualitativo
Teoría de la señalización

ABSTRACT

In spite of the relatively scarce use of textual analysis and sentiment analysis techniques in finance and accounting, they have great potential in accounting, both because of the volume of documents used for the communication of information and due to the growth in the use of digital tools and social media. In that regard, these techniques of analysis may help researchers to analyse hidden clues or look for additional information to that one observed through financial information, increasing the quantity and quality of the information traditionally used, and providing a new perspective of analysis. The aim of this study is to review the use of textual analysis and sentiment analysis in accounting. After presenting the concepts of textual analysis and sentiment analysis and expose their interest in accounting, we perform a review of the previous literature on the use of these techniques in finance and accounting and describe the main techniques of sentiment analysis, as well as the procedure to be followed for the use of this methodology. Finally, we suggest three lines of future research that may benefit from the use of textual and sentiment analysis.

©2021 ASEPUC. Published by EDITUM - Universidad de Murcia. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Análisis textual y del sentimiento en contabilidad

RESUMEN

A pesar del relativamente escaso uso de técnicas de análisis textual y de análisis del sentimiento en finanzas y contabilidad, éstas tienen un gran potencial en contabilidad, tanto por el elevado volumen de documentos utilizados para la comunicación de información financiera como por el crecimiento en el uso de herramientas digitales y medios de comunicación social. En este sentido, estas técnicas de análisis pueden ayudar a los investigadores a analizar pistas ocultas o buscar información adicional a la observada a través de los estados financieros, incrementando la cantidad y calidad de la información tradicionalmente utilizada, y proporcionando una nueva perspectiva de análisis. Por ello, el objetivo de este estudio es realizar una revisión del uso del análisis textual y del análisis del sentimiento en contabilidad. Tras presentar los conceptos de análisis textual y análisis del sentimiento y justificar teóricamente su papel en la investigación en contabilidad, llevamos a cabo una revisión de la literatura previa en el uso de estas técnicas en finanzas y contabilidad y describimos las principales técnicas de análisis del sentimiento, así como el procedimiento a seguir para el uso de esta metodología. Finalmente, sugerimos tres líneas de investigación futura que pueden beneficiarse del uso del análisis textual y del análisis del sentimiento.

©2021 ASEPUC. Publicado por EDITUM - Universidad de Murcia. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Textual analysis is a set of techniques to extract information from textual sources for its use in data analysis, business intelligence, or for research purposes, among others (Loughran & McDonald, 2016). It is a multidisciplinary field of study, which has reached a high level of development in several fields of knowledge, but it is in an embryonic state in finance and accounting (Kearney & Liu, 2014; Fisher et al., 2016). Among the techniques encompassed in textual analysis, we highlight sentiment analysis. Sentiment is defined as the level of polarity (positivity or negativity), as well as other sentiment dimensions (anxiety/calmness, optimism/pessimism), transmitted by the analysed text. Thus, sentiment analysis refers to the use of textual sentiment in order to identify and extract subjective information about the analysed text. The use of this methodology, which has been applied more intensely in other fields of knowledge, has been increasing, particularly in the field of behavioural finance.

In that regard, in spite of the relatively scarce use of textual analysis and sentiment analysis techniques in finance and accounting, they have great potential, both because of the volume of documents used for the communication of information, such as financial statements, earnings press releases, or corporate social responsibility reports, and due to the growth in the use of digital tools and social media. In fact, researchers have a wide range of possibilities for the application of these methodologies in our field of knowledge (Fisher et al., 2016; Loughran & McDonald, 2016).

We should note that textual documents, such as the MD&A report, include useful information that cannot be included in the financial statements, complementing them (Abrahamson & Amir, 1996; Bryan, 1997; Barron et al., 1999), and thus the use of textual analysis provides a better understanding of the quantitative information traditionally used¹, as well as a new perspective of analysis in, among other research areas (Li, 2010a; Kearney & Liu, 2014; Amani & Fadlalla, 2017): i) the study of corporate information disclosures, by examining whether the tone in the texts included in the documents may contain hidden clues about the actual situation of companies that is not explicitly shown by information from financial statements; ii) the analysis of the information from information intermediaries (auditors, financial analysts, credit rating agencies), which may contain additional information to that observed through the ratings and/or opinions shown in the reports; iii) sentiment analysis may provide additional relevant data about the evolution of the financial markets, by examining the contents of press news, and their relationship with other data used by analysts; and iv) finally, its application to the information from the Internet, a key aspect because of the importance of this information source and its global reach to the whole society. Sentiment analysis is especially interesting for its application to social media, because of the large number of messages disseminated and because it allows one to examine its influence on the markets, such as the stock price or the volume of transactions, as well as the effect that may have on other information sources, such as press news or financial analysts. We have to note that textual analysis may also be useful for practitioners; in that sense, it provides analysts with additional tools that complement the use of financial information, helping them in stock valuation, and providing alternative measures for the

investor sentiment; regarding the auditors, the textual analysis of non-financial information may help on the detection of accounting irregularities and fraudulent activities, or the prediction of bankruptcy.

For this reason, the aim of this study is to review the use of textual analysis and sentiment analysis in accounting. To do so, we carried out a search of articles about the use of these techniques in the top journals in Accounting (Accounting Review, Contemporary Accounting Research, Journal of Accounting Research, Journal of Accounting and Economics, Journal of Accounting and Public Policy, and Accounting Forum), as well as in accounting journals that examine the impact of the new technologies in accounting (Intelligent Systems in Accounting, Finance and Management, International Journal of Accounting Information Systems, Journal of Information Systems, and Journal of Emerging Technologies in Accounting). The period we examined runs from 2008 to 2019, a period that is explained by the relatively novelty of the textual analysis techniques, though we have included earlier references when we consider it advisable. Articles from these journals were selected based on the search of the keywords “textual analysis” and “sentiment analysis”. This search was complemented with the analysis of top journals in Finance (e.g. Journal of Finance or The Review of Financial Studies) and journals from other areas that regularly publish articles about accounting and corporate disclosures (e.g. Government Information Quarterly), as well as with the analysis of top journals in Computer Science and Artificial Intelligence, such as Computational Intelligence or Decision Support Systems, among others.

The main contributions of this paper are twofold: i) the first one is that it summarises the empirical studies in finance and accounting which have used textual and sentiment analysis. Although there are some papers that have reviewed previous literature (Kearney & Liu, 2014; Das, 2014; Fisher et al., 2016; Loughran & McDonald, 2016), they have focused on finance (Das, 2014; Kearney & Liu, 2014) or are more devoted to the explanation of the technical matters (Fisher et al., 2016; Loughran & McDonald, 2016). This paper tries to offer a more comprehensive view of textual and sentiment analysis in accounting, explaining the most relevant trends in their use; and ii) the second contribution is the proposal of new lines of research, linking traditional research in accounting (earnings quality, auditing) with the use of a relatively new set of techniques that may shed light on the usefulness of accounting information, which can help researchers in the application of textual and sentiment analysis in accounting.

The paper is structured as follows: First, after presenting the concepts of textual analysis and sentiment analysis, we perform a review of the previous literature in finance and accounting about the use of these techniques, taking into account the hypothesis that sentiment in text may have a signalling function on the traditional financial information. Secondly, we examine the main techniques of sentiment analysis, as well as the procedure to be followed for the use of this methodology. Finally, we suggest three lines of future research that may benefit from the use of sentiment analysis.

2. Textual analysis: Concept and literature review

2.1. Concept of textual analysis

The concepts of textual analysis, computational linguistics, natural language processing, or content analysis, refer to a set of techniques that lets researchers extract information from textual sources for its use in data analysis, business in-

¹Textual information contains detailed information about the numeric financial data, as well as additional non-financial information that may have relevance in the decision-making process of lenders and investors (Abrahamson & Amir, 1996).

telligence, or for research purposes, among others. These techniques can be considered a subset of qualitative analysis. Because of the nature of the analysis, it is an inherently multidisciplinary field of study, which relates different areas such as psychology (Noecker et al., 2013), computation sciences (Sudhahar et al., 2015), linguistics (Taboada, 2016), biomedical sciences (Cohen & Hunter, 2008) or neurosciences (Jorgensen, 2005), these being disciplines where textual analysis has reached a higher level of development. Textual analysis has also been used as an alternative methodology for literature reviews (Chakraborty et al., 2014; Hutchison et al., 2018).

Globally, textual analysis studies are encompassed in two trends (Loughran & McDonald, 2016): i) studies that examine the readability of text, i.e. those which measure the level of reader ability needed to understand the content of a text (Schroeder & Gibson, 1990; Li, 2008; Goel et al., 2010; Loughran & McDonald, 2014; Frankel et al., 2016; Asay et al., 2017; Bonsall et al., 2017; Dyer et al., 2017); and ii) studies that deal with the extraction of information from the text, either the search of keywords and targeted phrases (Loughran et al., 2009), topic modelling (Blei et al., 2003; Ball et al., 2015; Huang et al., 2018), the use of document similarity measures (Lang & Stice-Lawrence, 2015; Hoberg & Phillips, 2016), or sentiment analysis (Davis & Tama-Sweet, 2012; Hajek & Olej, 2013; Jegadeesh & Wu, 2013; Allee & DeAngelis, 2015).

With regard to the studies on readability, they are focused on the estimation of a measure that lets researchers assess the level of difficulty in the reading comprehension of the text (Loughran & McDonald, 2016). For example, the Fog Index (Li, 2008; Lo et al., 2017; Lim et al., 2018), similarly to other traditional readability measures, estimates the number of years of education needed to understand the text on a first reading. This measure is calculated based on the average sentence length and the percentage of complex words (words with more than two syllables; it is assumed that longer words are more difficult to understand).

Nevertheless, the use of traditional readability measures involves problems in accounting, because the use of long words, with are commonly understood in the field of accounting, do not necessarily involve greater complexity in the text (Loughran & McDonald, 2016). Moreover, similarly to earnings quality measures, the separation of the document complexity from the intrinsic complexity of the business is rather difficult (Leuz & Wysocki, 2016). For this reasons, Loughran & McDonald (2014, 2016) propose other alternative measures, such as the file size. Other readability measures take into account other factors such as the typographic style, the presence of non-textual elements like pictures, or the use of the passive voice (Schroeder & Gibson, 1990; Asay et al., 2018). On the other hand, the use of XBRL provides a structured context of data in order to facilitate the interpretation of the text by computers.

Regarding the search for keywords and targeted phrases, in spite of being one of the simplest textual analysis approaches, it has great potential because of its focus on a small range of words or phrases instead of a wide list of expressions, which can involve greater ambiguity (Loughran et al., 2009; Loughran & McDonald, 2016). With regard to topic modelling, it is a textual analysis technique that looks for the underlying topics in a set of documents, based on the correlations between the words in those documents (Blei et al., 2003; Huang et al., 2018). Regarding document similarity measures, they normally rely on a method labelled cosine similarity, for which, given two vectors of words (the docu-

ments), the normalised similarity value takes values from 0 to 1 (Brown & Tucker, 2011; Lang & Stice-Lawrence, 2015; Hoberg & Phillips, 2016). With regard to sentiment analysis, we develop in more detail both the concept and the techniques of analysis in Sections 3 and 4.

2.2. Textual analysis in finance and accounting

Although the use of textual analysis in accounting and finance is in its initial stage, it has great potential to be applied because of the significant volume of documentation which is used to disclose economic and financial information, such as financial statements, audit reports, corporate social reports, management reports, accounting standards, or analysts' reports, among others, which rely more on the value of textual, just nor numerical data (Amani & Fadlalla, 2017). Moreover, the exponential growth in the use of digital tools and social media by companies also increases significantly the volume of non-structured documents that are available on the Internet (Fisher et al., 2016). The online availability of press articles, conference calls², registers from governmental agencies like SEC, or text from social networks such as Facebook or Twitter, provide a large number of sources for the application of these techniques (Loughran & McDonald, 2016), including in accounting, since as suggested by Li (2010a: 143), "The textual information can provide a very useful context for understanding the financial data and testing interesting hypotheses."

We have to note that information from financial statements³ is rather concise, since it summarizes the financial position and performance in a few documents. Furthermore, although managers have some discretion when reporting financial information, it is rather limited as long as managers have to prepare this information according to accounting principles and rules. Although these limits to managers' discretionality are desirable as long as they help to mitigate earnings management and accounting manipulation, avoid an excessive level of optimism by managers, and help to enhance the comparability and reliability of financial information, these limits also hinder the ability of managers to communicate information that does not meet the reporting criteria.

In this regard, text in corporate disclosures helps to better understand the numbers from the financial statements (Abrahamson & Amir, 1996; Clatworthy & Jones, 2003), and lets the managers signal issues that could be unnoticed without the support of textual information from the Notes and the Discussion and Analysis Management Report. In that sense, the textual analysis of these sources can be especially useful, complementing the more traditional financial analysis of quantitative information from the financial statements. Among the potential sources for the textual analysis we have to highlight the Notes to the financial statements, the Discussion and Management Analysis Report, or the earning press releases

²A conference call is a teleconference or webcast in which a listed company reports the earnings of a certain period via the Internet, in order to communicate financial information immediately, broadly, and inexpensively to all investors (Bushee et al., 2003). Conference calls have been regulated in the US since the beginnings of this century (Bushee et al., 2004), and previous literature has shown that firms that regularly hold conference calls experience significant and sustained reductions in information asymmetry (Brown et al., 2004).

³An exception to this conciseness is the information provided in the Notes. Although the Notes are part of the financial statements and have an essential role to understand the rest of the financial statements, we consider this document separately because its narrative nature impede its use through traditional techniques (i.e. ratio analysis).

(Davis et al., 2012; Lang & Stice-Lawrence, 2015; Goel & Uzuner, 2016; Koo et al., 2017).

On the other hand, textual analysis can also be used to look for hidden cues in the corporates disclosures, in the sense that textual sources can contain information that does not support the perception from the financial statements, either because there is an intent to “sweeten” the actual financial situation and performance of the company, or because it is unintentionally unmasking deceiving quantitative information. Some examples of these hidden cues may be related to differences between quantitative and textual information, or the use of imagery, pleasantness, or ambiguity in the textual information (Humpherys et al., 2011). In that sense, textual analysis may help the information users to “read between lines”, in order to look for these hidden cues. This analysis can be made not only to detect potential fraudulent practices (Humpherys et al., 2011; Goel & Gangolly, 2012; Goel & Uzuner, 2016; Chen et al., 2017), but also to analyse the reports from the information intermediaries, such as credit scoring agencies, financial analysts, and auditors. Textual analysis can also be used to analyse unstructured data in an automated way, rather than manually, such as comment letters (Karim et al., 2019).

Among the papers that have carried out a literature review on the use of textual analysis in accounting and finance, we should highlight those of Das (2014), Fisher et al. (2016), and Loughran & McDonald (2016). Das (2014) reviews the academic literature that has used techniques of textual analysis in finance, and provides a user guide to those prepared to start out in textual analysis. Fisher et al. (2016) carry out a literature review about the use of natural language processing in accounting, auditing and finance, and they give some useful tips on its implementation. Finally, Loughran & McDonald (2016) explain the details on the use of the textual analysis, as well as giving some tips about its implementation.

In that regard, Table 1 summarises the main contributions on the use of textual analysis in finance and accounting, which have been grouped according to the information sources and the specific technique used. Therefore, studies are presented in four groups: i) those focused on the analysis of contents from corporate disclosures (Li, 2008; Loughran et al., 2009; Goel et al., 2010; Brown & Tucker, 2011; Davis et al., 2012; Hájek & Olej, 2013; Jegadeesh & Wu, 2013); ii) those which examine the information content from the analysts' reports and documents from other information intermediaries (Huang et al., 2014; Huang et al., 2018; Karim et al., 2019); iii) those which examine the impact of press news (García, 2013; Li et al., 2014a; Li et al., 2014b; Mo et al., 2016; Boudoukh et al., 2019; and iv) those which examine information from the Internet, especially that obtained from social media (Antweiler & Frank, 2004; Das & Chen, 2007; Bollen et al., 2011; Sprenger et al., 2014a; Sprenger et al., 2014b; Zheludev et al., 2014; Piñeiro-Chousa et al., 2017).

Among the studies in finance, the most common used sources have been press news, both general and specialised press (García, 2013; Li et al., 2014a; Li et al., 2014b; Mo et al., 2016), and the Internet (Das & Che, 2007; Zhang et al., 2011; Sprenger et al., 2014a; Siganos et al., 2017). Press news have traditionally been an important information source, and these texts are relevant to know the general conditions of the economy, the financial markets, and the industries and companies.

With regard to the studies in accounting which have used textual analysis, the most used sources are those related to corporate disclosures, such as the annual report (Li, 2008;

Goel et al., 2010; Goel & Gangolly, 2012; Lang & Stice-Lawrence, 2015), the 10-K report (Loughran & McDonald, 2011; Jegadeesh & Wu, 2013), the Management Discussion and Analysis (MD&A) Report (Loughran et al., 2009; Li et al., 2010; Davis and Tama-Sweet, 2012; Ball et al., 2015; Goel & Uzuner, 2016; Buehlmaier & Whited, 2018), earnings press releases (Henry, 2006; Davis et al., 2012; Koo et al., 2017), or conference calls (Allee & DeAngelis, 2015). Some of these studies have used sentiment analysis as the textual analysis technique (Loughran & McDonald, 2011; Davis et al., 2012; Davis & Tama-Sweet, 2012; Hájek & Olej, 2013; Allee & DeAngelis, 2015; Goel & Uzuner, 2016), and thus they are described in Section 3.

Corporate disclosures are a natural source for textual analysis, since they are official communications from insiders who have access to information and a better knowledge of the firm than outsiders. The writing style and the tone of these texts may contain useful information about the future performance of the company and the data on the financial statements, and it is particularly useful to examine the role of qualitative information on individual performance and stock pricing, so it is a factor to be taken into account among the fundamentals which are analysed in event studies. Nevertheless, a limitation in the use of this information source is related to its timing (quarterly or annual information), as well as the potential bias introduced in the text by managers with expressions that try to deceive investors about the true business situation.

In that regard, several studies have looked for the detection of fraudulent activities by analysing the text from the annual reports, the 10-K reports or the MD&A reports (Cecchini et al., 2010; Goel et al., 2010; Humpherys et al., 2011; Goel & Gangolly, 2012; Goel & Uzuner, 2016; Chen et al., 2017). These studies are based on the premise that text from these reports have hidden cues, and they are focused on samples of companies involved in civil lawsuits or accounting and auditing enforcement releases.

Other studies that examine the MD&A reports have the aim of predicting fraud or bankruptcy (Cecchini et al., 2010; Goel et al., 2010; Mayew et al., 2015). Cecchini et al. (2010) create a dictionary of differentiating concepts. This dictionary can differentiate fraudulent companies from honest companies on 75% of occasions, and healthy companies from bankrupt companies 80% of the time. After combining (soft) textual data with quantitative data, which have been commonly used in the literature about fraud and bankruptcy prediction, they increase the accuracy of their predictions up to 81.97% for fraud prediction and 83.87% for bankruptcy prediction.

In this line, Goel et al. (2010) examine both the verbal content and the presentation style of qualitative information from annual reports, with the intention of exploring linguistic features that may help to distinguish the fraudulent reports from the honest ones. They find systematic differences in the presentation and communication style of the two groups of reports, and they find evidence that the examination of the linguistic features is a useful tool for fraud detection, being able to improve the accuracy of their initial predictions from 56.75% when using a “bag of words” approach to 89.51% after the incorporation of linguistic features.

In a contemporary study, Humpherys et al. (2011) examine the use of deceptive language in financial fraud by managers. To do so, the authors look for linguistic cues in publicly available corporate disclosures. They find that fraudulent disclosures use more words, activation language, imagery, pleasantness, group references, and less lexical diversity than non-fraudulent disclosures. These results suggest that writers of

Table 1
Textual analysis in finance and accounting

Information Source	Technique of textual analysis	Authors
Corporate disclosures	Content analysis	Humpherys et al. (2011), Goel et al. (2010)
	Analysis of writing and presentation style	Goel & Gangolly (2012)
	Sentiment analysis	Cecchini et al. (2010), Cho et al. (2010), Li (2010), Loughran & McDonald (2011), Davis et al. (2012), Davis & Tama-Sweet (2012), Hájek & Olej (2013), Jegadeesh & Wu (2013), Barkemeyer et al. (2014), Allee & DeAngelis (2015), Mayew et al., 2015, Goel & Uzuner (2016), Chen et al. (2017)
	Readability	Li (2008), Goel et al. (2010), Frankel et al. (2016), Asay et al. (2017); Bonsall et al. (2017) Dyer et al. (2017), Lim et al. (2018), Melloni et al. (2017)
	Document similarity	Brown & Tucker (2011), Lang & Stice-Lawrence (2015)
	Targeted phrases	Loughran et al. (2009)
Analysts' reports and other information intermediaries' documents	Topic modelling	Ball et al. (2015)
	Sentiment analysis.	
	Analysis of writing and presentation style	Twedt & Rees (2012), Huang et al. (2014)
Internet and social media	Topic modelling	Huang et al. (2018), Karim et al. (2019)
	Sentiment analysis	Antweiler & Frank (2004), Das & Chen (2007), Bollen et al. (2011), Zhang et al.(2011), Chen et al. (2014), Kim & Kim (2014), Sprenger et al. (2014a),Sprenger et al. (2014a), Sprenger et al. (2014b), Sprenger et al. (2014b),Zheludev et al. (2014),Nguyen et al. (2015), Souza et al. (2016), Sul et al. (2017), Piñeiro-Chousa et al. (2017), Siganos et al. (2017)
Press news	Sentiment analysis	Twedt & Rees (2012), García (2013), Li et al. (2014a), Li et al. (2014b), Malo et al. (2014), Mo et al. (2016), Zhang et al. (2016), Boudoukh et al. (2019)
	Document similarity	Tetlock (2011)

fraudulent disclosures try to appear more credible, although they communicate less content. The findings support the use of linguistic analysis by auditors for the assessment of risk fraud, and to signal questionable financial disclosures. Mayew et al. (2015) find that both management's opinion about going concern reported in the MD&A and the linguistic tone of the MD&A help to predict whether a firm will go to bankruptcy. In a more recent study, Chen et al. (2017) develop a fraud detection method for narratives in annual reports, combining natural language processing, queen genetic algorithm and support vector machine.

A second stream of research that uses corporate disclosures as the textual source is that related to ethics and corporate social responsibility (Loughran et al., 2009; Cho et al., 2010; Barkemeyer et al., 2014; Melloni et al., 2017). Loughran et al. (2009) examine the occurrence of terms related to ethics in 10-K reports and they find evidence that companies which use more "ethical" terms are more likely to be "sin" companies, i.e. companies to be object of lawsuits and to score poorly on measures of corporate governance, what suggests that the use of ethical terms in their reports is linked with the purpose of deceiving the public. On the other hand, Barkemeyer et al. (2014) examine CEO statements in sustainability reports, and they find evidence that companies with poorer performance in sustainability terms employ more ambiguous language with the intention of disguising their bad performance, as commonly occurred in financial reporting. Finally, Melloni et al. (2017) examine a sample of early Integrated Reports (IR) adopters and they find that, in presence of weak financial performance, the IR tends to be less readable and more optimistic.

The third stream of research using corporate disclosure examines their information content and its explicative and predictive ability to predict future financial statements (Li, 2010b) or bankruptcy (Cecchini et al., 2010; Hájek & Olej, 2013), to explain financial performance (Li, 2008; Davis et al., 2012), for pricing purposes (Elrod, 2009; Ball et al.,

2015), or to estimate the level of accruals (Frankel et al., 2016). Li (2010b) examines the information content of the section of the MD&A report referring to financial statements forecasting. This study finds evidence that companies with better performance, lower accruals and higher readability have more positive forecasts.

Li (2008) examines the association between the readability of the annual report and the performance, as well as earnings persistence. He measures report readability using the Fog Index and document length, and finds evidence that the annual reports of companies with lower earnings are more difficult to read, and that companies with more persistent earnings disclosure provide more readable reports. On the other hand, Davis et al. (2012) examine the information content of earnings press releases using sentiment analysis tools, which we explain in more detail in Section 3.

With regard to the use of textual analysis for valuation purposes, Ball et al. (2015) use a topic modelling approach on the MD&A reports with the aim of obtaining information about the companies for which the value relevance of financial statement is relatively low because of business changes. They find evidence that relevant discussions can be grouped into specific topics that explain the nature of the business change, such as investment strategies, securities issuance, or financial constraints. Finally, Frankel et al. (2016) use textual analysis to assess the contents of the qualitative disclosures, estimating accruals based on information from MD&A reports and relating them with actually reported accruals. They find evidence that estimated accruals explain a significant portion of actual accruals, and they help to identify the most persistent accruals. Furthermore, they find evidence that the explanatory power of estimated accruals is higher for more readable reports.

With regard to studies that apply textual analysis to obtain information from analysts' reports, we have to note that these reports contain references to both corporate disclosures and press news, and they may have information that has been pre-

pared for insiders to be communicated to investors and other market participants. Moreover, while news usually talk about past facts, analyst reports are more focused on information about estimations and expectations, so the analysis of these texts may have higher predictive power. As stated by [Twedt & Rees \(2012\)](#), financial analysts have an essential role in the assessment of financial information and the dissemination of their analyses to investors, which explains the interest in their reports by both investors and academics.

To date, only four studies have used reports from analysts and other information intermediaries as the textual analysis source ([Twedt & Rees, 2012](#); [Huang et al., 2014](#); [Huang et al., 2018](#); [Karim et al., 2019](#)). [Twedt & Rees \(2012\)](#) examine whether the level of detail (complexity, length, visual resources) and the tone of the reports affect the market response. It is expected that a higher level of detail may reflect the efforts of analysts to prepare the report, and thus the usefulness of their estimations, while the tone may signal the underlying opinion of the analysts about the company, which can be used to assess whether the analysts have conflicts of interest, and how they may affect their estimations. In line with these hypotheses, the results show that the tone contains incremental information content to the reports' predictions and recommendations, and the reports' complexity helps to explain changes in the market response to analysts' recommendations.

[Huang et al. \(2014\)](#) use sentiment analysis techniques, as well as writing and presentation style, with the aim of assessing the information content of the analysts' reports. Specifically, they examine the textual opinions from analysts following S&P500 companies for the period 1995-2008, and find evidence that textual analysis provides additional information to the quantitative measures from the reports, showing that the market reacts more intensely to favourable (unfavourable) quantitative information when textual opinion is more positive (negative). Furthermore, they observe that negative opinions have a higher weight, suggesting that analysts play an important role in the dissemination of bad news.

In a more recent study, [Huang et al. \(2018\)](#) use Topic Modelling techniques on analysts' reports and corporate disclosures, with the aim of examining the role of information intermediaries. The study shows that analysts do not merely examine the issues dealt in the conference calls, but they also discuss exclusive additional issues. Moreover, they find evidence that investors value new information provided by analysts when managers have incentives to hide relevant information. In sum, the study exposes the key role of financial analysts as information intermediaries.

Finally, [Karim et al. \(2019\)](#) examine the comment letters on two exposure drafts proposed by the FASB. The authors use a Latent Dirichlet allocation to automatically analyse the contents of these comment letters, what enables them to identify topics and detect shift in focus of the letters responding to the exposure drafts.

3. Sentiment analysis: Concept and use in finance and accounting

Among the textual analysis techniques, we find sentiment analysis, which relies on the identification and extraction of subjective information from a text based on the level of polarity (positivity or negativity) transmitted by it ([Kearney & Liu, 2014](#)). It is considered that sentiment appears in several ways in the human discourse: public speeches, reports, news, blogs, and other forms of written, spoken and visual communication ([Kearney & Liu, 2014](#)). As with research in textual

analysis, some fields other than accounting and finance have reached a high level of development in the application of sentiment analysis. For example, in psychology, there is evidence of a correlation between the sentiment expressed in social networks and the political positions of parties and politicians ([Tumasjan et al., 2011](#)). In marketing, some papers have examined the association between the sentiment expressed in opinions and the demand for the products ([Duan et al., 2008a, 2008b](#); [Ye et al., 2009](#)). In linguistics, the association on the use of certain categories of words (verbs, nouns, adjectives and adverbs) has been examined with the tone of the text, as well as with the aim of detecting fake news, based on the sentiment expressed in the text ([Taboada, 2016](#)).

In finance, the use of sentiment analysis is especially linked to behavioural finance ([Li et al., 2014a](#); [Mo et al., 2016](#)). In this area, researchers have made great efforts to understand how sentiment affect the decision making process for individuals, institutions and the market. In a general sense, sentiment can be divided into two streams: i) investor sentiment, as the combination of beliefs about the future evolution of cash flows and investment risks, not justified by known and/or rational facts ([Baker & Wugler, 2007](#)); and ii) textual sentiment, referring to the level of polarity of the text, which can be expressed in terms of positivity/negativity, but also in other dimensions, such as strong/weak, or active/passive ([Goel & Uzuner, 2016](#); [Loughran & McDonald, 2016](#)). Therefore, investor sentiment captures subjective judgements and characteristics of the investors' behaviour, while textual sentiment includes, in addition to investor sentiment, other more objective reflections about the conditions of companies, institutions and markets ([Kearney & Liu, 2014](#)).

Sentiment analysis in accounting is theoretically supported by the fact that the behaviour of both preparers and users of financial information may be far from rational, affecting their decision making process; as shown by prior literature, presenting information in positive terms results in more favourable evaluations than when information is presented in negative terms ([Levin et al., 1998](#)), what also applies to corporate information ([Davis et al., 2012](#)). Moreover, when considering other more unregulated alternative sources, such as the Internet, this "emotional" behaviour is even more important, as documented by the "social network" effect examined by [Saxton & Wang \(2014\)](#). Therefore, we have to note that the decision making process of the information users is not only affected by the amount and quality of the information, but also by the sentiment transmitted by the information, an issue that is also known by the information preparers, who may try to affect the users' decisions through the way they communicate the information. For these reasons, the sentiment analysis applied to accounting information provides accounting researchers with a powerful tool to examine the behaviour of both users and preparers of accounting information.

As we stated in the previous section, most of the prior literature about the use of sentiment analysis in finance has used news and the Internet as the information source, linking these sources to several capital market measures, such as returns, volatility, or volume of transactions ([Li et al., 2014a](#); [Mo et al., 2016](#)). Among the studies that have used news, [García \(2013\)](#) examines the effect of sentiment on the stock price for the period between 1905 and 2005. Using the fraction of positive and negative words in two financial columns of the New York Times as a measure for sentiment, he finds evidence that the predictive ability of news is concentrated in recessions. On the other hand, [Li et al. \(2014a\)](#) find evidence that information from news affects activity in the

capital market, and public sentiment may cause emotional fluctuations in investors, affecting the decision making process; moreover, this impact changes depending on the firm characteristics and the news content.

In another study, Li et al. (2014b) use several sentiment measures, based on Harvard & Loughran-McDonald dictionaries, and they assess the association between news and the market in Hong Kong. Finally, Mo et al. (2016) examine the feedback effect between news sentiment and market returns. Specifically, they find evidence that the effect of sentiment on news has a delay of five days, while the delay for the effect of returns on sentiment is only one day. These results suggest that news sentiment drives activity in markets and investment decisions, triggering involuntary responses that are expressed with higher press coverage and affecting news sentiment. Boudoukh et al. (2019) show that firm-level public news is a meaningful component of stock return variance, and they identify from news stories relevant public information tied to specific firm events.

The other most used source of textual sentiment in finance is the Internet, especially social networks (Das & Chen, 2007; Bollen et al., 2011; Zhang et al., 2011; Chen et al., 2014; Sprenger et al., 2014a; Sprenger et al., 2014b; Piñeiro-Chousa et al., 2017; Siganos et al., 2017). The disclosure of information about finance and accounting on the Internet is a potentially useful source for sentiment analysis, since there is a wide and diverse audience who interacts in an active (writing) or passive (reading) way on the Internet and in social networks. Sentiment analysis applied to the Internet may help to detect the market sentiment, the existence of opportunistic behaviours, and the reaction of Internet users to other information sources (Das & Chen, 2007).

Nevertheless, since information on the Internet tends to be open and unregulated, the quantification of sentiment on the Internet is likely more “noisy” and has greater bias than when using other sources, and it can have little information incremental to published news. Moreover, a high proportion of messages are written by noisy traders or poorly informed investors, so they are susceptible to particular opinions and sentiments, and information may tend to be less precise and reliable. On the other hand, the pre-processing of the information from the Internet is costlier than when using corporate disclosure and press news, because people tend to write in a less precise, clear and formal way on the Internet, and the meaning of the text may be more ambiguous.

However, although sentiment on the Internet is potentially noisy, it is a potential tool for extracting sentiment from small investors. The analysis of social media is especially interesting, not only in the sphere of for-profit organisations (Jiang et al., 2009; Bonsón et al., 2011), but also public administrations, such as city councils (Bonsón et al., 2012; Alcaide Muñoz et al., 2014; Bonsón et al., 2015; Gandía et al., 2016), and non-profit organisations (Saxton & Wang, 2014; Gálvez-Rodríguez et al., 2016). The amount of opinions, information, and tweets in finance provide the opportunity to examine the influence of these sources on stock pricing or the volume of transactions. Furthermore, their analysis lets researchers examine the differential effect of social media on the more conventional contents of the Internet (Luo et al., 2013), as well as the effect that social media can have on other information sources, such as analysts’ reports (Sprenger et al., 2014b). Recent studies show the pre-eminence of social media on the information about financial markets (Souza et al., 2016).

Studies using the Internet and social networks examine the relationship between sentiment in social media and the

reaction in the capital markets. Although most of studies have used Twitter (Bollen et al., 2011; Zhang et al., 2011; Sprenger et al., 2014a, 2014b), other studies have used financial platforms, such as Yahoo Finance (Antweiler & Frank, 2004; Das & Chen, 2007; Nguyen et al., 2015), Seeking Alpha (Chen et al., 2014), or Stocktwits (Piñeiro-Chousa et al., 2017). The use of general social networks like Twitter captures a more generic sentiment, while the sentiment captured through the use of financial social media like Stocktwits is more specific (Debreceeny et al., 2019), so a higher link between the sentiment of these social media and capital markets is expected.

With regard to the use of corporate disclosures, the inclusion of sentiment analysis techniques provides a new perspective, complementing the quantitative information that has been traditionally used (Li, 2010a; Kearney & Liu, 2014). Sentiment analysis of corporate disclosures is of interest to the extent that the tone in documents may contain hidden cues about the actual situation of companies, which cannot be explicitly shown by financial statements, such as financial constraints (Hájek & Olej, 2013) or accounting irregularities that may lead to fraud (Goel & Uzuner, 2016). Therefore, sentiment analysis can contribute to research in accounting because of the signalling effect of language and tone used in the documents (Cho et al., 2010; Davis et al., 2012). Since the study of Loughran and McDonald (2011), who examine 10-K reports with the aim of testing the effectiveness of generic lists, as opposed to the used of dictionaries of specific terms⁴, the number of studies using sentiment analysis techniques on corporate disclosures has increased (Amernic et al., 2010; Davis et al., 2012; Davis & Tama-Sweet, 2012; Hájek & Olej, 2013; Jegadeesh & Wu, 2013; Barkemeyer et al., 2014; Alee & DeAngelis, 2015; Goel & Uzuner, 2016).

A commonly used source from corporate disclosures is earnings press releases (Davis et al., 2012; Davis & Tama-Sweet, 2012). Earnings press releases generally precede MD&A reports and are followed by financial media, analysts and investors, but regulation about their format and contents is rather limited, allowing managers more flexibility (Davis & Tama-Sweet, 2012). Davis et al. (2012) examine the information content of earnings press releases using sentiment analysis. Specifically, they examine the association between sentiment in earnings press releases and the market response. The authors find empirical evidence that an optimistic speech is positively associated with future returns on assets, and it involves a positive response by the market. Results suggest that language has the purpose of, both directly and implicitly, signalling the expectations of managers about future performance. On the other hand, Davis & Tama-Sweet (2012) consider that the impact on capital markets is higher after the earnings press releases than after the publication of the annual reports, and they theorise that managers will strategically choose the language depending on the communication channel to be used. They examine the narrative disclosure in MD&A reports and earnings press releases, and they find evidence that companies exhibit more optimistic language in earnings press releases.

With regard to the use of annual reports, they have been used for risk prediction (Hájek & Olej, 2013; Tsai & Wang, 2017) and fraud detection (Goel & Uzuner, 2016), and the association between textual sentiment found in them and the market reaction has been examined (Jegadeesh & Wu, 2013; Hájek, 2018). Regarding risk prediction, Hájek & Olej (2013) examine the effect of sentiment on future financial difficulties, assessing the sentiment in the annual reports of US

⁴Use of lists and other techniques of analysis is explained in Section 4.

companies. Combining the use of financial indicators with sentiment analysis, the authors find evidence that information from textual sentiment improves significantly the accuracy of the classifiers. On the other hand, Tsai & Wang (2017) see the association between risk prediction and textual information (considered to be soft information) as a complement to information from financial statements (hard information). To do this, they order companies based on their risk levels through ranking techniques, and they examine information from MD&A reports about the future evolution of the company.

Regarding the use of sentiment in annual reports for fraud detection, Goel & Uzuner (2016) examine the relationship between sentiment in MD&A reports and fraud. The authors measure the sentiment in the dimensions of polarity, subjectivity, and intensity, in order to investigate whether fraudulent and non-fraudulent reports differ in such dimensions. Results show that fraudulent reports contain three times more positive sentiment, and four times more negative sentiment, as compared to honest reports, what suggests that the use of sentiment is more pronounced in fraudulent reports, which have also a higher proportion of subjective contents as compared to honest reports.

With regard to the association between sentiment in annual reports and the market response, Jegadeesh & Wu (2013) examine the relationship between sentiment in 10-K reports and market response, and they find evidence of a significant association. Hájek (2018) combines the use of financial indicators with sentiment measures obtained from the annual reports in order to improve the accuracy of the models to predict stock price.

Regarding the analysis of conference calls, Allee & DeAngelis (2015) examine the tone dispersion in order to assess whether the narrative structure provides cues about voluntary disclosures by managers, as well as their effect on users. They find evidence that tone dispersion is associated with current and future performance, reporting decisions, and managers' incentives to mislead perceptions on information. Furthermore, they find evidence that tone dispersion is associated with analysts' and investors' responses to narratives used in conference calls.

The last source from corporate disclosure is that related to transparency and sustainability reports. Barkemeyer et al. (2014) examine whether corporate sustainability reports are a precise representation of the performance in terms of sustainability, based on the sentiment analysis of CEO statements in these reports and in financial reports. In spite of the increasing standardisation in sustainability reporting because of GRI, their analysis shows that rhetoric used in sustainability reports is more related to impression management than to accountability, and thus it is in line with reporting used to communicate financial performance rather than to communicate sustainability policies and results.

4. Sentiment analysis techniques

4.1. Content analysis

The analysis of textual sentiment as a research technique requires the use of specific methods to extract sentiment from text. The two most commonly used approaches in literature are the use of dictionaries or lexicons (Kothari et al., 2009; Loughran & McDonald, 2011; Davis & Tama-Sweet, 2012; Twedt & Rees, 2012; Loughran & McDonald, 2016; Goel & Uzuner, 2016) and machine learning (Antweiler & Frank,

2004; Das & Chen, 2007; Li, 2010b; Hájek & Olej, 2013; Huang et al., 2014).

4.1.1. Dictionary approach

The dictionary-based approach, also known as "bag of words", uses a mapping algorithm by which a program reads the text and classifies words, phrases and sentences in predefined categories (Li, 2010a). Documents are considered bags of words (Buehlmaier & Whited, 2018), which are examined through sentiment lexicons or dictionaries containing words with an associated semantic orientation. By matching words in the text with predefined lists of words labelled with a specific sentiment, researchers can determine the semantic orientation of the text (Goel & Uzuner, 2016). Researchers using this approach must take into account two issues: 1) the dictionary to be used, and 2) the weighting of the words in the total list.

With regard to the dictionary, the literature in accounting and finance has mainly used four lists: The Henry (2008) list, the Harvard General Inquirer (GI), Diction, and the Loughran & McDonald (2011) list. Other lists that have been recently used are SentiStrength (Zheludev et al., 2014) and SentiwordNet (Mo et al., 2016). The Henry list, specifically created for the analysis of financial text, was originally used to examine earnings press releases, and it has also been used for the analysis of conference calls (Price et al., 2012; Davis et al., 2012). Its main weakness is the limited number of words on the list, which means that a significant part of the text is not examined.

Harvard General Inquirer (Tetlock, 2007, Kothari et al., 2009; Loughran & McDonald, 2011; Twedt & Rees, 2012; Li et al., 2014b) is a software program that maps text files based on several dictionary categories, and it applies an algorithm able to assign words to these categories. It has been one of the most used because of its earlier availability, and it has been extensively used for the analysis of press news (Tetlock, 2007), as well as for corporate disclosures (Kothari et al., 2009) and initial prospectuses (Hanley & Hoberg, 2010).

Another dictionary that has been used is Diction (Elrod, 2009; Amernic et al., 2010; Cho et al., 2010; Davis et al., 2012; Goel & Uzuner, 2016; Melloni et al., 2017). In addition to including standard lists of words, it allows for the creation of ad hoc lists. The output of the processed text includes statistics about the total number of words, the number of characters, the average size of words, and the number of different words. It also provides counts for special characters and high frequency words.

We have to note that the use of Harvard General Inquirer and Diction presents several limitations. Since they are general dictionaries and have not been specifically created for financial vocabulary, some words can be classified as negative in a general context when they are not so in finance, either because these words have a technical meaning (such as tax, cost or liability) or because they are linked to certain activities (such as crude, cancer, mine or death) (Loughran & McDonald, 2011, 2016). As a consequence of their general nature, Li (2010b) finds evidence that tone classification based on these dictionaries does not provide enough accuracy. On the other hand, Loughran and McDonald (2011) find that 73.8% of negative words in Harvard General Inquirer are not negative in a financial context. Errors in the classification may go beyond merely adding noise, and may unintentionally create latent measures of other companies' attributes, such as size or industry.

In order to mitigate these limitations, Loughran & McDon-

ald (2011) created a new dictionary with six lists of words that capture six sentiment categories (negative, positive, uncertainty, litigious, strong modal and weak modal). This dictionary has two advantages over the previous dictionaries: 1) lists are relatively comprehensive; and 2) they have been specifically created for the analysis of financial communication. Kearney & Liu (2014) state that it is the most used dictionary in finance recently (Chen et al., 2014; García, 2013; Hájek & Olej, 2013; Jegadeesh & Wu, 2013; Li et al., 2014b).

With regard to the weighting of words, most studies use proportional weighting, treating each word in the list with the same importance. Loughran & McDonald (2011) use two schemes: a proportional one and another one which is inversely proportional to the word frequency in the document. Jegadeesh & Wu (2013) consider that there is no reason to expect that less frequent words should have higher weight, so they use a scheme in which weighting depends on the market reaction to the specific word in the past. On the other hand, another issue to be considered is the lemmatisation process (Mo et al., 2016), i.e. the conversion of different words with the same root into a single word (e.g. the conversion of “rising”, “risen” and “rises” in “rise”).

The use of dictionaries to measure tone have several advantages: 1) once the dictionary has been selected, there is no subjectivity by the researcher; 2) since the count is carried out in programs, researchers can work with large samples; and 3) since there are publicly available dictionaries, researchers can replicate the work of previous studies in a simple way. Nevertheless, despite these advantages and their relatively generalised use, this approach has several limitations.

First, the previous polarity of a word, defined as the sentiment the word transmits without a context (like “beautiful” or “horrible”), may be different from the global semantic orientation of a sentence; one of the challenges for textual analysis is going beyond considering words as independent entities, taking into account the interactions between them (Loughran & McDonald, 2016). Moreover, we should note that, while in other fields, like product or film reviews, sentiments are expressed as combinations of adjectives and adverbs, financial sentiment is more related to the expected directions of events from an investor perspective (e.g. “it is expected the profit will increase”). In order to mitigate these limitations, Malo et al. (2014) develop a model (Linearized Phrase Structure, LPS) to detect semantic orientations in pieces of economic-financial text, which complements the use of dictionaries.

A second limitation is that certain publications, especially those from social media, may contain grammatical or lexical errors, which can make sentiment analysis difficult. With the aim of mitigating this limitation, Zheludev et al. (2014) use SentiStrength, a sentiment classification system adapted to the often-incorrect nature of text in social media. However, it is not prepared to identify elements of human speech, such as sarcasm⁵, and thus it may erroneously classify text based on the used words.

4.1.2. Machine Learning approach

Machine Learning is often considered part of artificial intelligence (Fisher et al., 2016), and it was initially used in maths and computational sciences. Its use in textual analysis is based on the application of statistical techniques to infer

the contents of documents and to classify them, according to the correlation between the frequency of some words and the reference document (Li, 2010b). Essentially, the process has three stages (Kearney & Liu, 2014):

- (1) Selection of part of the text, which is manually examined. This portion is considered the “training set”, classifying each word with attributes, such as “positive”, “negative”, or other sentiment dimensions to be analysed, such as “strong/weak” or “active/passive”.
- (2) Application of a selection of algorithms on the training set; these algorithms “learn” the sentiment classification rules.
- (3) Application of the sentiment classification rules to the whole text.

Therefore, Machine Learning involves the use of one or more algorithms that work on the training set and prepare a model containing its statistics, which are later applied to the whole corpus in order to estimate an index of textual sentiment (Kearney & Liu, 2014). Among the studies that have used Machine Learning are those of Antweiler & Frank (2004), Das & Chen (2007), Li (2010b), and Huang et al. (2014). Some of the applications that have been used are Rainbow (Antweiler & Frank, 2004; Das & Chen, 2007) and Reuters NewsScope Sentiment Engine.

Rainbow is a software program developed by McCallum (1996), which permits several alternative methods of classification, such as Naïve Bayes, Term Frequency – Inverse Document Frequency, or K-nearest neighbour. With regard to NewsScope Sentiment Engine, it lets researchers do the pre-processing of the text, which is used to lexically identify words as nouns, verbs, adjectives or adverbs, after which the classification of the sentiment is made through neural networks. Regarding the algorithms used in Machine Learning, some of the most commonly used are Support Vector Machines Artificial Neural Networks, or Bayesian Networks.

Compared to the dictionary approach, dictionaries are easier to apply for analysts because of the availability of programs, which explains their wide use in prior literature. Limitations related to the use of general dictionaries can be mitigated by the use of specific lists. Furthermore, the implementation of Machine Learning is costlier and requires more time because the training set must be manually classified. On the other hand, in order to guarantee the highest quality in the manual stage, the selection of the people that will classify the tests must be very strict.

In favour of the Machine Learning approach, it can be used when there are not specific dictionaries for the language or document to be analysed (Li, 2010b). Secondly, as opposed to the dictionary approach, Machine Learning does take the context of a sentence into consideration. In addition, since the training set is manually codified, it can be used to test the effectiveness of the algorithm. We should note that studies using Machine Learning usually show a higher accuracy rate than studies using dictionaries (Li et al., 2010; Huang et al., 2014).

4.2. Measurement of textual sentiment

Once the sentiment has been extracted from the text, the estimation of sentiment measures to compare documents or to use them for other research purposes is relatively straightforward. Nevertheless, because of the differences among the methods of sentiment analysis, the characteristics of the estimated measure may be different (Kearney & Liu, 2014).

⁵While the dictionary approach is not suitable to it, the Machine Learning approach has been used to identify sarcasm, becoming a relevant topic in natural language processing.

In dictionary-based studies, the most common sentiment measure is the percentage of words of a specific category compared to the total number of words in the text (Kothari et al., 2009; Ferguson et al., 2015; Chen et al., 2014), or the standardised percentage (Tetlock et al., 2008). Standardisation is needed when the raw frequency is not constant, e.g. when the writing style depends on the author. In the case of standardised values, we should note that, although the raw percentage will always be equal to or higher than zero, normalised percentages may be negative.

Another alternative measure is the number of positive words less the number of negative words, divided by the sum of both (Twedt & Rees, 2012). The main advantage of this measure is that it classifies a text as relatively positive or negative, depending on the specific weight of the categories. Finally, other studies have used principal components analysis (Tetlock, 2007; Doran et al., 2012; Price et al., 2012). The main advantage of this approach is that the sentiment to be extracted is not decided a priori, but it depends on the weight of the total variation among all the categories, thus showing the global style of the text. Nevertheless, this approach is not suitable when researchers want to focus on a specific sentiment.

With regard to the studies that use Machine Learning, the estimated measure is based on the global classification of the texts. For example, Das & Chen (2007) develop a sentiment index estimated as the cumulated classification of messages, depending on the sentiment of each message. Li (2010b) assigns the value of 1 to positive sentences, 0 to neutral sentences, and -1 to negative sentences. For each document, the global sentiment is calculated as the average score of the individual sentences. This process is also followed by Huang et al. (2014) on the analysts' reports. By way of contrast, Hájek & Olej (2013) calculate the frequency of net positive words as the difference between positive and negative terms.

Regardless of whether dictionaries or Machine Learning techniques are used, an important issue to be taken into account is the existence of a neutral category, either to detect the documents encompassed in this category to consider them in the analysis, or to exclude them (Koppel & Schler, 2006; Taboada et al., 2011).

5. Future of textual analysis and sentiment analysis in accounting

Once researchers have decided the information source to be analysed and the sentiment analysis technique to be employed, they estimate a sentiment measure, which will be used to test the hypotheses with regard to the association between the sentiment measure and the variables of interest. In that regard, the sentiment measure can be used as a variable in the traditionally used models in accounting research. In this section, we state some of the lines of research that could benefit from the use of textual and sentiment analysis.

5.1. Sentiment in corporate disclosures and earnings quality: looking for hidden cues

As explained in Sections 2.2 and 3, some studies in accounting have used textual analysis and sentiment analysis to assess corporate disclosures. The linguistic style and tone in the texts may have useful information, containing hidden cues about the actual situation of companies that are not explicitly shown in the financial statements (Elrod, 2009; Hájek & Olej, 2013; Goel & Uzuner, 2016). On the other hand, similarly to the use of earnings management with opportunistic

purposes (García Osma et al., 2005; Lo, 2008), management may use qualitative disclosures to mislead the judgement of investors. For these reasons, textual analysis and sentiment analysis have a great potential to be used on corporate disclosures. As explained in Section 2.2, previous literature has used textual analysis in the detection of fraudulent activities, the analysis of ethics and corporate social responsibility issues, and the analysis of the information content of corporate disclosures.

To date, however, there is scarce research about the association between earnings quality and text content in corporate disclosures; only Frankel et al. (2016) have examined the relationship between the content of corporate disclosures and accruals. Accordingly, we should note the role of textual sentiment as a signalling element (Davis et al., 2012), so sentiment analysis can be used to examine the association between the language used in corporate disclosures and earnings management (Lo et al., 2017), or to compare the optimism transmitted by qualitative corporate disclosures with accounting conservatism from financial statements. Moreover, future research should examine whether there are significant differences between earnings quality measures and sentiment in corporate disclosures and, if there do exist differences, the economic consequences of these differences.

On the other hand, determinants of the observed sentiment in corporate disclosures should be examined, such as the language used in the disclosures, the characteristics of the firm (size, industry, leverage, performance), the personal characteristics of the preparer of the disclosures (country of origin, age, academic education or genre), or the motivations for managers to express more/less sentiment in disclosures, in line with the motivations for earnings management.

5.2. Sentiment analysis and social media: organisation's sentiment vs. followers' sentiment

A second line is related to the analysis of the social media, both the organisations' information and the followers' messages about the organisations. Previous literature has shown the importance of disclosures by organisations, both for-profit and non-profit organisations, because of the reduction of information asymmetries and the uncertainty faced by them (Francis et al., 2005). In for-profit organisations, information disclosures decrease information risk and enhance their financing conditions (Francis et al., 2008; Bharath et al., 2008). Given the restrictions on financial information (with issues related to reporting and accounting GAAP), companies may choose to disclose additional non-financial information through other alternative channels that complement financial information. Accordingly, the Internet offers a very flexible channel to corporations; furthermore, the wide use of social media among investors and consumers makes its use recommendable for the reporting purposes of companies (Debrecey et al., 2019).

On the other hand, with regard to non-profit organisations (NPOs), there are several differences that affect both the nature and the reach of the disclosed information. First, unlike investors in the corporate setting, NPO donors do not expect to receive an economic return for their investment, but its utility is defined in "social" terms, through the "mission-related performance" (Saxton et al., 2014); for this reason, traditional corporate disclosures, focused on financial and economic characteristics, may not be useful for donors, because of both a lack of knowledge and the inability of financial information to measure performance in terms other than the economic ones. Moreover, we have to note that the

nature of the information affects the communication channel. Given the wide variety of users and the relevance of qualitative and voluntary information for NPOs, the Internet (websites, but especially social media) is an ideal communication channel for them (Gandía, 2011; Saxton et al., 2014).

We should note that marketing research has shown that the demand for a product is affected by the Internet users' opinions (Duan et al., 2008a, 2008b; Ye et al., 2009) and previous literature in accounting and finance also shows that Internet disclosures about a company affect the pricing of their stock (Tetlock, 2007; García, 2013; Ferguson et al., 2015). In that regard, both streams show that the information disclosed by a third party about a company may have a higher economic impact than the information disclosed by the company. Furthermore, although the Internet and the social media have characteristics that may help the disclosure of information for both companies and NPOs, we have to note that, because of both the mostly qualitative nature of information and the less specialised users' profile, social media users exhibit more emotional and less rational behaviour than traditional information users (Lovejoy & Saxton, 2012; Guo & Saxton, 2014), so the effects of users' sentiment in social media for investment and purchase decisions is an open question; hence, the use of sentiment analysis may be a potential tool to examine the information in social media, not only that reported by organisations, but also for third parties, as well as to examine how social media users react to these information outflows, and their economic consequences.

5.3. Audit reports: reading between lines

A third line is that related to information intermediaries, including financial analysts, auditors or credit scoring agencies. Textual analysis may provide a new perspective on the role of these agents, both in the theoretical framework of the agency theory, for which information intermediaries have an important function in the assessment of the accounting information, and in the signalling function that these intermediaries may have with regard to financial information (Huguet & Gandía, 2014, 2016). In that regard, textual analysis of qualitative documents may explain scorings and/or opinions given by information intermediaries, either credit scores, analysts' recommendations or audit opinions, as well as finding hidden cues that may disagree with the intermediary's public position (Does the credit scoring agency rely on their own estimations? Is the auditor masking qualifications?).

Textual analysis of the auditors' report is especially interesting because of the importance of audit research in accounting. Although audit reports are highly standardised, auditors have had some discretion with the inclusion of matter paragraphs. Furthermore, the changes in the audit report under the International Standards of Auditing (ISAs), with the introduction of Key Audit Matters (KAM), as well as matters requiring significant auditor attention (e.g. higher assessed risks and significant risks, areas of significant management judgement and estimation uncertainty, or significant transactions or events) have created a natural setting to test whether the new audit reports have become more informative than the previous ones. In that regard, in order to avoid a limited comparison between the previous and the new audit reports, textual analysis should be used to examine in depth the differences between both models. Finally, sentiment analysis may create audit-based sentiment measures that can be included in audit quality research.

6. Conclusions

The volume of textual, qualitative documents used as a complement to financial statements, such as MD&A reports, earnings press releases, or corporate social responsibility reports, and the growth in the use of digital tools and social media offer a new perspective of analysis for accounting research through the use of textual and sentiment analysis. These tools can help researchers to examine whether the tone in the texts of the corporate disclosures contain hidden clues that are not explicitly shown by numerical information, if there is additional information in the information intermediaries reports that remains unobserved through their ratings and/or opinions, and whether the sentiment extracted from the texts try to influence on users' decision making process.

Therefore, the aim of this study is to review the use of textual analysis in accounting. After an introduction of the concepts of textual and sentiment analysis, and the exposition of the reasons that explain the usefulness and convenience of these techniques, we perform a review of the previous literature in finance and accounting on the use of textual and sentiment analysis, and present the main techniques of analysis, as well as the procedure to be followed when applying this methodology. Finally, we propose three lines of future research for which textual and sentiment analysis can be especially useful: i) sentiment in corporate disclosures and quality of the financial disclosure; ii) corporate and users' sentiment on the Internet and social media; and iii) textual analysis of the audit reports.

This paper contributes to previous literature in two ways: first, this is the first paper that summarises the empirical studies in finance and accounting which have used textual and sentiment analysis, complementing the previous ones which have focused on finance or on technical matters, by providing a motivation for the use of these techniques, and by explaining the most relevant trends in the use of textual analysis and sentiment analysis in accounting; secondly, the paper proposes new lines of research considering traditional research in accounting through the use of a new set of techniques that can contribute to shed light on the usefulness of accounting when complemented to textual, qualitative information.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Conflict of interests

The authors declare no conflict of interests.

References

- Abrahamson, E., & Amir, E. (1996). The information content of the president's letter to shareholders. *Journal of Business, Finance and Accounting*, 23(8), 1157-1181. <https://doi.org/10.1111/j.1468-5957.1996.tb01163.x>.
- Alcaide Muñoz, L., Rodríguez Bolívar, M.P., & Sánchez, R.G. (2014). Estudio cuantitativo de la investigación en transparencia informativa, participación ciudadana y prestación de servicios públicos mediante la implementación del e-gobierno. *Revista de Contabilidad – Spanish*

- Accounting Review, 17(2), 130-142. <https://doi.org/10.1016/j.rcsar.2014.05.001>
- Allee, K.D., & DeAngelis, M.D. (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>
- Amani, F. A., & Fadlalla, A. M. (2017). Data mining applications in accounting: A review of the literature and organizing framework. *International Journal of Accounting Information Systems*, 24, 32-58. <https://doi.org/10.1016/j.accinf.2016.12.004>
- Amernic, J., Craig, R., & Tourish, D. (2010). Measuring and assessing tone at the top using annual report CEO letters. The Institute of Chartered Accountants of Scotland. [https://researchportal.port.ac.uk/portal/en/publications/measuring%2dand%2dassessing%2dtone%2dat%2dthe%2dtop%2dusing%2dannual%2dreport%2dceo%2dletters\(5f009fa3%2d76fe%2d441f%2d91c3%2dc8535273d71f\).html](https://researchportal.port.ac.uk/portal/en/publications/measuring%2dand%2dassessing%2dtone%2dat%2dthe%2dtop%2dusing%2dannual%2dreport%2dceo%2dletters(5f009fa3%2d76fe%2d441f%2d91c3%2dc8535273d71f).html)
- Antweiler, W., & Frank, M.Z. (2004). Is all that talk just noise? The information content of Internet stock message boards. *The Journal of Finance*, 59(3), 1259-1294. <https://doi.org/10.1111/j.1540-6261.2004.00662.x>
- Asay, H.S., Elliott, W.B., & Rennekamp, K. (2017). Disclosure readability and the sensitivity of investors' valuation judgments to outside information. *The Accounting Review*, 92(4), 1-25. <https://doi.org/10.2308/accr-51570>
- Asay, H.S., Libby, R., & Rennekamp, K. (2018). Firm performance, reporting goals, and language choices in narrative disclosures. *Journal of Accounting and Economics*, 65(2-3), 380-398. <https://doi.org/10.1016/j.jacceco.2018.02.002>
- Baker, M., & Wugler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21(2), 129-151. <https://doi.org/10.3386/w13189>
- Ball, C., Hoberg, G., & Maksimovic, V. (2015). *Disclosure, business change and earnings quality*. Working Paper. Available at SSRN: <https://ssrn.com/abstract=2260371>.
- Barkemeyer, R., Comyns, B., Figge, F., & Napolitano, G. (2014). CEO statements in sustainability reports: Substantive information or background noise? *Accounting Forum*, 38(4), 241-257. <https://doi.org/10.1016/j.accfor.2014.07.002>
- Barron, E.E., Kile, C.O., O'keefe, T.B. (1999). MD&A quality as measured by the SEC and analysts' earnings forecasts. *Contemporary Accounting Research*, 16(1), 75-109. <https://doi.org/10.1111/j.1911-3846.1999.tb00575.x>
- Bharath, S.T., Sunder, J., & Sunder, S.V. (2008). Accounting quality and debt contracting. *The Accounting Review*, 83(1), 1-28. <https://doi.org/10.2139/ssrn.591342>
- Blei, D.M., Ng, A. Y., & Jordan, M.I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1-8. <https://doi.org/10.1016/j.jocs.2010.12.007>
- 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>
- Bonsón, E., & Flores, F. (2011). Social media and corporate dialogue: The response of the global financial institutions. *Online Information Review*, 35(1), 34-49. <https://doi.org/10.1108/14684521111113579>
- Bonsón, E., Royo, S., & Ratkai, M. (2015). Citizen's engagement on local governments' Facebook sites. An empirical analysis: The impact of different media and content types in Western Europe. *Government Information Quarterly*, 32(1), 52-62. <https://doi.org/10.1016/j.giq.2014.11.001>
- Bonsón, E., Torres, L., Royo, S., & Flores, F. (2012). Local e-government 2.0: Social media and corporate transparency in municipalities. *Government Information Quarterly*, 29(2), 123-132. <https://doi.org/10.1016/j.giq.2011.10.001>
- Boudoukh, J., Feldman, R., Kogan, S., & Richardson, M. (2019). Information, trading and volatility: Evidence from firm-specific news. *The Review of Financial Studies*, 32(3), 992-1033. <https://doi.org/10.1093/rfs/hhy083>
- Brown, S., Hillegeist, S.A., & Lo, K. (2004). Conference calls and information asymmetry. *Journal of Accounting and Economics*, 37(3), 343-366. <https://doi.org/10.1016/j.jacceco.2004.02.001>
- Brown, S.V., & Tucker, J.W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, 49(2), 309-346. <https://doi.org/10.1111/j.1475-679X.2010.00396.x>
- Bryan, S.H. (1997). Incremental information content of required disclosures contained in Management Discussion and Analysis. *The Accounting Review*, 72(2), 285-301.
- Buehlmaier, M.M.M., & Whited, T.M. (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., Matsumoto, D.A., & Miller, G.S. (2003). Open versus closed conference calls: The determinants and effects of broadening access to disclosure. *Journal of Accounting and Economics*, 34, 149-180. [https://doi.org/10.1016/S0165-4101\(02\)00073-3](https://doi.org/10.1016/S0165-4101(02)00073-3)
- Bushee, B.J., Matsumoto, D.A., & Miller, G.S. (2004). Managerial and investor responses to disclosure regulation: The case of Reg FD and conference calls. *Journal of Accounting Research*, 79(3), 617-643. <https://doi.org/10.2139/ssrn.310233>
- Cecchini, M., Aytug, H., Koehler, G.J., & Pathak, P. (2010). Detecting management fraud in public companies. *Management Science*, 56(7), 1146-1160. <https://doi.org/10.1287/mnsc.1100.1174>
- Chakraborty, V., Chiu, V., & Vasarhelyi, M. (2014). Automatic classification of accounting literature. *International Journal of Accounting Information Systems*, 15(2), 122-148. <https://doi.org/10.1016/j.accinf.2014.01.001>
- Chen, H., De, P., Hu, J., & Wang, B.H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies*, 27(5), 1367-1403. <https://doi.org/10.1093/rfs/hhu001>
- Chen, Y.J., Wu, C.H., Chen, Y.M., Li, H.Y., & Chen, H.K. (2017). Enhancement of fraud detection for narratives in annual reports. *International Journal of Accounting Information Systems*, 26(1), 32-45. <https://doi.org/10.1016/j.accinf.2017.06.004>
- Cho, C.H., Roberts, R.W., & Patten, D.M. (2010). The language of US corporate environmental disclosure. *Accounting, Organizations and Society*, 35(4), 431-443. <https://doi.org/10.1016/j.aos.2009.10.002>
- Clatworthy, M., & Jones, M.J. (2003). Financial reporting of good news and bad news: Evidence from accounting narratives. *Accounting and Business Research*, 33(3), 171-185. <https://doi.org/10.1080/00014788.2003.9729645>
- Cohen, K.B., & Hunter, L. (2008). Getting Started in Text Mining. *PLoS Computational Biology*. 4(1), e20. <https://doi.org/10.1371/journal.pcbi.1000020>

- [//doi.org/10.1371/journal.pcbi.0040020](https://doi.org/10.1371/journal.pcbi.0040020)
- Das, S., & Chen, M. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9), 1375-1388. <https://doi.org/10.1287/mnsc.1070.0704>
- Das, S. (2014). Text and context: Language analytics in Finance. *Foundations and Trends in Finance*, 8(3), 145-261. <https://doi.org/10.1561/05000000045>
- Davis, A.K., Piger, J.M., & Sedor, L.M. (2012). Beyond the numbers: Measuring the information content of earning press release language. *Contemporary Accounting Research*, 29(3), 845-868. <https://doi.org/10.1111/j.1911-3846.2011.01130.x>
- Davis, A.K., & 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>
- Debreceny, R.S., Want, T., & Zhou, M. (2019). Research in social media: Data sources and methodologies. *Journal of Information Systems*, 33(1), 1-28. <https://doi.org/10.2308/isisys-51984>
- Doran, J.S., Peterson, D.R., & Price, S.M. (2012). Earnings conference call content and stock price: The case of REITs. *Journal of Real Estate Finance and Economics*, 45(2), 402-434. <https://doi.org/10.1007/s11146-010-9266-z>
- Duan, W., Gu, B., & Whinston, A.B. (2008a). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems*, 45(4), 1007-106. <https://doi.org/10.1016/j.dss.2008.04.001>
- Duan, W., Gu, B., & Whinston, A.B. (2008b). The dynamics of online word-of-mouth and product sales – An empirical investigation of the movie industry. *Journal of Retailing*, 84(2), 233-242. <https://doi.org/10.1016/j.jretai.2008.04.005>
- 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>
- Elrod, G.B. (2009). *Is there predictive value in the words managers use? A key word analysis of the annual reports' Management Discussion and Analysis (Doctoral dissertation)*. University of Texas at Arlington.
- Ferguson, N.J., Philip, D., Lam, H.Y.T., & Guo, J.M. (2015). Media content and stock returns: The predictive power of press. *Multinational Finance Journal*, 19(1), 1-31. <https://doi.org/10.17578/19-1-1>
- Fisher, I.E., Garnsey, M.R., & Hughes, M.E. (2016). Natural Language Processing in accounting, auditing, and finance: A synthesis of the literature with a roadmap for future research. *Intelligent Systems in Accounting, Finance and Management*, 23(3), 157-214. <https://doi.org/10.1002/isaf.1386>
- Francis, J.R., Khurana, I.K., & Pereira, R. (2005). Disclosure incentives and effects on cost of capital around the world. *The Accounting Review*, 80(4), 1125-1162. <https://doi.org/10.2308/accr.2005.80.4.1125>
- Francis, J., Dhananjay, N., & Olsson, P. (2008). Voluntary disclosure, earnings quality, and cost of capital. *Journal of Accounting Research*, 46(1), 53-99. <https://doi.org/10.1111/j.1475-679X.2008.00267.x>
- Frankel, R., Jennings, J., & Lee, J. (2016). Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, 62(2-3), 209-227. <https://doi.org/10.1016/j.jacceco.2016.07.003>
- Gálvez-Rodríguez, M.M., Caba-Pérez, C., & López-Godoy, M. (2016). Drivers of Twitter as a strategic communication tool for non-profit organizations. *Internet Research*, 26(5), 1052-1071. <https://doi.org/10.1108/IntR-07-2014-0188>
- Gandía, J.L. (2011). Internet disclosure by non-profit organizations: Empirical evidence of nongovernmental organizations for development in Spain. *Nonprofit and Voluntary Sector Quarterly*, 40(1), 57-78. <https://doi.org/10.1177/0899764009343782>
- Gandía, J.L., & Huguet, D. (2018). Differences in audit pricing between voluntary and mandatory audits. *Academia Revista Latinoamericana de Administración*, 31(2), 336-359. <https://doi.org/10.1108/ARLA-01-2016-0007>
- Gandía, J.L., Marrahí, L., & Huguet, D. (2016). Digital transparency and Web 2.0 in Spanish city councils. *Government Information Quarterly*, 33(1), 28-39. <https://doi.org/10.1016/j.giq.2015.12.004>
- García, D. (2013). Sentiment during recessions. *The Journal of Finance*, 68 (3), 1267-1300. <https://doi.org/10.1111/jofi.12027>
- García Osma, B., Gill de Albornoz, B., & Gisbert, A. (2005). La investigación sobre earnings management (Research on earnings management). *Spanish Journal of Finance and Accounting*, 34(127), 1001-1033. <https://doi.org/10.1080/02102412.2005.10779570>
- Goel, S., Gangolly, J., Faerman, S.R., & Uzuner, O. (2010). Can linguistic predictors detect fraudulent financial filings? *Journal of Emerging Technologies in Accounting*, 7(1), 25-46. <https://doi.org/10.2308/jeta.2010.7.1.25>
- Goel, S., & Gangolly, J. (2012). Beyond the numbers: Mining the annual report for hidden cues indicative of financial statement fraud. *Intelligent Systems in Accounting, Finance and Management*, 19(2), 75-89. <https://doi.org/10.1002/isaf.1326>
- Goel, S., & Uzuner, O. (2016). Do sentiments matter in fraud detection? Estimating semantic orientation of annual reports. *Intelligent Systems in Accounting, Finance and Management*, 23(3), 215-239. <https://doi.org/10.1002/isaf.1392>
- Guo, C., & Saxton, G. (2014). Tweeting social change: How social media are changing Nonprofit advocacy. *Nonprofit and Voluntary Sector Quarterly*, 43(1), 57-79. <https://doi.org/10.1177/0899764012471585>
- Hájek, P., & Olej, V. (2013). Evaluating sentiment in annual reports for financial distress prediction using Neural Networks and Support Vector Machines. *Engineering Applications of Neural Networks*, pp 1-10, in *International Conference on Engineering Applications of Neural Networks*. <https://doi.org/10.1007/978-3-642-41016-1%5f1>
- 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>
- Hales, J., Kuang, X.I., & Venkataraman, S. (2011). Who believes the hype? An experimental examination of how language affects investor judgments. *Journal of Accounting Research*, 49(1), 223-255. <https://doi.org/10.1111/j.1475-679X.2010.00394.x>
- Hanley, K.W., & Hoberg, G. (2010). The information content of IPO prospectuses. *Review of Financial Studies*, 23, 2821-2864. <https://doi.org/10.1093/rfs/hhq024>
- Henry, E. (2006). Market reaction to verbal components of earnings press releases: Event study using a predictive algorithm. *Journal of Emerging Technologies in Accounting*,

- 3(1), 1-19. <https://doi.org/10.2308/jeta.2006.3.1.1>
- Henry, E. (2008). Are investors influenced by how earnings press releases are written? *The Journal of Business Communication*, 45(4), 363-407. <https://doi.org/10.1177/0021943608319388>
- Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124(5), 1423-1465. <https://doi.org/10.3386/w15991>
- Huang, A.H., Zang, A.Y., & 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, A.H., Lehavy, R., Zang, A.Y., & Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science*, 64(6), 2833-2855. <https://doi.org/10.1287/mnsc.2017.2751>
- Huguet, D. and Gandía, J.L. (2014). Cost of debt capital and audit in Spanish SMEs. *Spanish Journal of Finance and Accounting / Revista Española de Financiación y Contabilidad*, 43(3), 266-289. <https://doi.org/10.1080/02102412.2014.942154>
- Huguet, D. and Gandía, J.L. (2016). Audit and earnings management in Spanish SMEs. *Business Research Quarterly*, 19(3), 171-187. <https://doi.org/10.1016/j.brq.2015.12.001>
- Humpherys, S.L., Moffitt, K.C., Burns, M.B., Burgoon, J.K., & Felix, W.F. (2011). Identification of fraudulent financial statements using linguistic credibility analysis. *Decision Support Systems*, 50(3), 585-594. <https://doi.org/10.1016/j.dss.2010.08.009>
- Hutchison, P.D., Daigle, R.J., & George, B. (2018). Application of latent semantic analysis in AIS academic research. *International Journal of Accounting Information Systems*, 31(1), 83-96. <https://doi.org/10.1016/j.accinf.2018.09.003>
- Jegadeesh, N., & Wu, D. (2013). Word power: A new approach for content analysis. *Journal of Financial Economics*, 110(3), 712-729. <https://doi.org/10.1016/j.jfineco.2013.08.018>
- Jiang, Y., Raghupathi, V., & Raghupathi, W. (2009). Content and design of corporate governance web sites. *Information Systems Management*, 26(1), 13-27. <https://doi.org/10.1080/10580530802384704>
- Jorgensen, P. (2005). Incorporating context in text analysis by interactive activation with competition artificial neural networks. *Information Processing and Management: An International Journal*, 41(5), 1081-1099. <https://doi.org/10.1016/j.ipm.2004.10.003>
- Karim, K.E., Lim, K.J., Pinsker, R.E., & Zhu, H. (2019). Using linguistics to mine unstructured data from FASB exposure drafts. *Journal of Information Systems*, 33(1), 67-83. <https://doi.org/10.2308/isys-51928>
- 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, S.H., & Kim, D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. *Journal of Economic Behavior and Organization*, 107(B), 708-729. <https://doi.org/10.1016/j.jebo.2014.04.015>
- Koo, D.S., Wu, J.J., & Yeung, P.E. (2017). Earnings attribution and information transfers. *Contemporary Accounting Research*, 34(3), 1547-1579. <https://doi.org/10.1111/1911-3846.12308>
- Koppel, M., & Schler, J. (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2), 100-109. <https://doi.org/10.1111/j.1467-8640.2006.00276.x>
- Kothari, S.P., Li, X., & Short, J.E., (2009). The effect of disclosures by management, analyst, and business press on cost of capital, return volatility, and analyst forecast: A study using content analysis. *The Accounting Review*, 84(5), 1639-1670. <https://doi.org/10.2308/accr.2009.84.5.1639>
- 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>
- Leuz, C., & Wysocki, P.D. (2016). The economics of disclosure and financial reporting regulation: Evidence and suggestions for future research. *Journal of Accounting Research*, 54(2), 525-622. <https://doi.org/10.1111/1475-679X.12115>
- Levin, I.P., Schneider, S.L., & Gaeth, G.J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, 76(2), 149-188. <https://doi.org/10.1006/obhd.1998.2804>
- Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), 221-247. <https://doi.org/10.1016/j.jacceco.2008.02.003>
- Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, 29, 143-165.
- 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, Q., Wang, T., Li, P., Liu, L., Gong, Q., & Chen, Y. (2014a). The effect of news and public mood on stock movements. *Information Sciences*, 278, 826-840. <https://doi.org/10.1016/j.ins.2014.03.096>
- Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014b). New impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69, 14-23. <https://doi.org/10.1016/j.knosys.2014.04.022>
- Lim, E.K.Y., Chalmers, K., & Hanlon, D. (2018). The influence of business strategy on annual report readability. *Journal of Accounting and Public Policy*, 37(1), 65-81. <https://doi.org/10.1016/j.jaccpubpol.2018.01.003>
- Lo, K. (2008). Earnings management and earnings quality. *Journal of Accounting and Economics*, 45(2-3), 350-357. <https://doi.org/10.1016/j.jaccpubpol.2018.01.003>
- Lo, K., Ramos, E., & Rogo, R. (2017). Earnings management and annual report readability. *Journal of Accounting and Economics*, 63(1), 1-25. <https://doi.org/10.1016/j.jacceco.2016.09.002>
- Loughran, T., McDonald, B., & Yun, H. (2009). A wolf in sheep's clothing: The use of ethics-related terms in 10-K reports. *Journal of Business Ethics*, 89(Supplement), 39-49. <https://doi.org/10.1007/s10551-008-9910-1>
- 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. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, 69(4),

- 1643-1671. <https://doi.org/10.1111/jofi.12162>
- 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>
- Lovejoy, K., & Saxton, G.D. (2012). Information, community and action: How non-profit organizations use social media. *Journal of Computer-Mediated Communication*, 17(3), 337-353. <https://doi.org/10.1111/j.1083-6101.2012.01576.x>
- Luo, X., Zhang, J., & Duan, W. (2013). Social media and firm equity value. *Information Systems Research*, 24(1), 146-163. <https://doi.org/10.2139/ssrn.2162167>
- Malo, P., Sinha, A., Takala, P., Korhonen, P., & Wallenius, J. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782-796. <https://doi.org/10.1002/asi.23062>
- Mayew, W.J., Sethuraman, M., & Venkatachalam, M. (2015). MD&A disclosure and the firm's ability to continue as a going concern. *The Accounting Review*, 90(4), 1621-1651. <https://doi.org/10.2139/ssrn.2272463>
- McCallum, A. (1996). Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. *Working paper: School of Computer Science, Carnegie-Mellon University*.
- Melloni, G., Caglio, A., & Perego, P. (2017). Saying more with less? Disclosure conciseness, completeness and balance in Integrated Reports. *Journal of Accounting and Public Policy*, 36(3), 220-238. <https://doi.org/10.1016/j.jaccpubpol.2017.03.001>
- Mo, S.Y.K., Liu, A., & Yand, S.Y. (2016). New sentiment to market impact and its feedback effect. *Environment Systems and Decisions*, 36(2), 158-166. <https://doi.org/10.1007/s10669-016-9590-9>
- Nguyen, T.H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24), 9603-9611. <https://doi.org/10.1016/j.eswa.2015.07.052>
- Noecker, J., Ryan, M., & Juola, P. (2013). Psychological profiling through textual analysis. *Literary and Linguistic Computing*, 28(3), 382-387. <https://doi.org/10.1093/lit/fqs070>
- Piñeiro-Chousa, Vizcaíno-González, M., & Pérez-Pico, A.M. (2017). Influence of social media over the stock market. *Psychology and Marketing*, 34(1), 101-108. <https://doi.org/10.1002/mar.20976>
- Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B.A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking and Finance*, 36(4), 992-1011. <https://doi.org/10.1016/j.jbankfin.2011.10.013>
- Saxton, G.D., & Wang, L. (2014). The social network effect: The determinants of giving through social media. *Nonprofit and Voluntary Sector Quarterly*, 43(5), 850-868. <https://doi.org/10.1177/0899764013485159>
- Siganos, A., Vagenas-Nanos, E., & Verwijmeren, P. (2017). Facebook's daily sentiment and international stock markets. *Journal of Economic Behavior and Organization*, 107(B), 730-743. <https://doi.org/10.1016/j.jebo.2014.06.004>
- Souza, T., Kolchyna, O., Treleaven, P.C., & Aste, T. (2016). *Twitter sentiment analysis applied to Finance: a case study in the retail industry*. Handbook of Sentiment Analysis in Finance. Mitra, G. and Yu, X. (Eds.). (2016). ISBN 1910571571.
- Sprenger, T.O., Sandner, P.G., Tumasjan, A., & Welpe, I. (2014). News or noise? Using Twitter to identify and understand company-specific news flow. *Journal of Business Finance and Accounting*, 41(7-8), 791-830. <https://doi.org/10.1111/jbfa.12086>
- Sprenger, T.O., Tumasjan, A., Sandner, P.G., & Welpe, I.M. (2014). Tweets and trades: The information content of stock microblogs. *European Financial Management*, 20(5), 926-957. <https://doi.org/10.1111/j.1468-036X.2013.12007.x>
- Schroeder, N., & Gibson, C. (1990). Readability of Management's Discussion and Analysis. *Accounting Horizons*, 4(4), 78-87.
- Sudhahar, S., Veltri, G.A., & Cristianini, N. (2015). Automated analysis of the US presidential elections using Big Data and network analysis. *Big Data and Society*, 2(1), 1-28. <https://doi.org/10.1177/2053951715572916>
- Sul, H.K., Dennis, A.R., & Yuan, L. (2017). Trading on Twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, 48(3), 454-488. <https://doi.org/10.1111/dec.12229>
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307. <https://doi.org/10.1162/COL1%5fa%5f00049>
- Taboada, M. (2016). Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2, 325-347. <https://doi.org/10.1146/annurev-linguistics-011415-040518>
- Tetlock, P.C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139-1168. <https://doi.org/10.2139/ssrn.685145>
- Tetlock, P.C., 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>
- Tetlock, P.C. (2011). All the news that's fit to reprint: Do investors react to stale information? *The Review of Financial Studies*, 24(5), 1481-1512. <https://doi.org/10.1093/rfs/hhq141>
- Tsai, M.F., & Wang, C.J. (2017). On the risk prediction and analysis of soft information in finance reports. *European Journal of Operational Research*, 257(1), 243-250. <https://doi.org/10.1016/j.ejor.2016.06.069>
- Tumasjan, A., Sprenger, T.O., Sandner, P.G., & Welpe, I.M. (2011). Election forecasts with Twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4), 402-418. <https://doi.org/10.1177/0894439310386557>
- Twedt, B., & Rees, L. (2012). Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy*, 31(1), 1-21. <https://doi.org/10.1016/j.jaccpubpol.2011.10.010>
- Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1), 180-182. <https://doi.org/10.1016/j.ijhm.2008.06.011>
- Zhang, X., Fuehres, H., & Gloor, P.A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia - Social and Behavioural Sciences*, 26, 55-62. <https://doi.org/10.1016/j.sbspro.2011.10.562>
- Zhang, J.L., Härdle, W.K., Chen, C.Y., & Bommers, E. (2016). Distillation of news flow into analysis of stock reactions.

Journal of Business and Economic Statistics, 34(4), 547-563. <https://doi.org/10.1080/07350015.2015.1110525>
Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? *Scientific Reports*, 4. <https://doi.org/10.1038/srep04213>