The International Journal of Digital Accounting Research
Vol. 17, 2017, pp. 1-32
ISSN: 2340-5058

A comparison of content analysis usage and text mining in CSR corporate disclosure

Selena Aureli. Bologna University. Italy selena.aureli@unibo.it

Abstract. This paper investigates content analysis and text mining as two different research techniques largely used by scholars to perform the text analysis of company social and environmental reports. Its aim is to demonstrate that these techniques are not so irreconcilable as one might suppose, but can be applied to the same problem being addressed in certain circumstances. The paper starts with a presentation of the two research techniques, providing information about their origin, diffusion and fields of application in academia. It examines them with respect to the assumption each one holds about the nature of knowledge and continues with a discussion of the elements that display the differences and similarities between them. Subsequently, the paper presents an empirical application of the two techniques to the same research problem: the identification of possible changes in the amount of company disclosure after an industrial disaster. The focus is on a company's communication of economic, social and environmental goals and impacts usually included in sustainability reports. The two techniques are applied to the same set of company reports published by four large multinationals that went through the industrial disaster characterised by strong negative social and environmental consequences. Results from the trend analysis obtained by the application of the two techniques indicate that they are not such irreconcilable methods, but they may lead to different conclusions about a company's behaviour in trying to restore its corporate reputation damaged by the disaster. Thus, the two techniques should not be used to crosscheck results, although they provide similar output data.

Keywords: Text mining, Content analysis, Legitimacy theory, Sustainability reporting, Disaster, crisis, Research methods

DOI: 10.4192/1577-8517-v17_1

Submitted September 2016
Accepted December 2016

1. INTRODUCTION

In the last decade, an increasing number of companies have been involved in the preparation and disclosure of sustainability reports (KPMG, 2013; 2015), which include information on economic, social and environmental dimensions as the major measures of the corporate sustainability quality - the so-called 'triple bottom line' (Elkington, 1998). Depending on the country examined, this reporting activity may be a voluntary practice, a result of legislation or part of a code of practice (Gray et al., 1995) and can be informed according to different frameworks and standards like UN Global Compact Principles, OECD Guidelines for Multinational Enterprises, GRI guidelines, ISO 26000, AA1000 and SA88000.

As an alternative to sustainability reports, companies may divulge information about their social and environmental goals, actions and related consequences through other documents like the integrated report, citizenship report, social report or even the annual report. At any rate, regardless of the form and type of report used, this type of disclosure usually aims to inform both internal and external stakeholders (Farneti and Guthrie, 2009), demonstrating transparency and accountability for company actions (Bellringer et al., 2011), as well as engage a dialogue with stakeholders and governments (Bonsón and Bednárová, 2014), legitimise company actions and attract investors (Bellringer et al., 2011; Adams and McNicholas, 2007). In other terms, as suggested by the stakeholder theory, political economy theory and the legitimacy theory (Gray et al., 1995; Adams, 2002), stakeholders represent the reason for disclosure (due to their increasing relevance and the pressure they put on corporations) and the addressee of this communication that companies try to educate, engage or manipulate (in the worst case).

The use of company disclosure to manage a company's relationships with stakeholders is particularly important in the case of industrial disasters. In fact, after such a negative event, managers may attempt to change stakeholders' perceptions and restore the company's legitimacy through an increase or decrease in the quantity of communication (Aerts and Cormier, 2009; Summerhays and De Villiers, 2012). In light of the legitimacy theory, managers activate communication strategies in order to acquire, maintain or re-acquire the

legitimacy of the business in society (Adams et al., 1998; Deegan, 2002; Bebbington et al., 2008; Michelon, 2011; Suchman, 1995; Cho and Patten, 2007; Bonilla-Priego et al., 2014). However, it is not possible to formulate a precise hypothesis about the direction of change in the amount of the information disclosed because past studies are contradictory and inconclusive. We might expect an increase in the amount of information produced (Cho, 2009; Islam and Deegan, 2010; Summerhays and De Villiers, 2012), however, a reduction in corporate disclosure is also possible if the company prefers an avoidance strategy that aims to distract public attention from the negative event (De Villiers and Van Staden, 2006; O'Dwyer, 2002).

Accounting researchers interested in exploring what is communicated to external subjects and which topics are most discussed in reports need adequate tools for data collection and data analysis. However, as suggested by Gray and Milne (2015), there is no superior research approach, method or technique. The researcher should identify the most appropriate technique available that follows the logical and methodological framework defined (Bryman and Bell, 2015).

Within business and accounting studies, the most common technique used to analyse economic, social and environmental information is content analysis (Krippendorff, 2004), which is *"a technique for gathering and analysing the content of text. The content refers to words, meanings, pictures, symbols, ideas, themes or any message that can be communicated"* (Neuman, 2003, p. 219). Content is coded into various categories or concepts depending on selected criteria (Weber, 1988), and coding can be manual or computer-aided.

This technique has been widely utilised with reference to social and environmental disclosures (Milne and Adler, 1999; Guthrie and Abeysekera, 2006; Beck et al., 2010) in order to identify the topics covered in sustainability reports (Holcomb et al., 2007; Guthrie and Farneti, 2008), to rank companies that disclose sustainability information creating quantitative scales and assigning scores (Clarkson et al., 2008; Michelon, 2011; Bonsón and Bednárová, 2015), to verify if sustainability information and disclosed indexes match the items proposed by international reporting standards like the GRI guidelines (Roca and Searcy, 2012; Tewari and Dave, 2012) and to describe company reactions to

industrial disasters and other legitimacy issues (Walden and Schwartz, 1997; De Villiers and Van Staden, 2006; Cho and Patten, 2007; Cho, 2009; O'Donovan, 2002; Summerhays and De Villiers, 2012).

Recently, text mining has also been used to analyse trends and patterns in sustainability reporting (Aureli et al., 2016) and identify which aspects are the most disclosed in company reports (Modapothala et al., 2010; Te Liew et al., 2014). Text mining is a computer assisted technique that is equipped with the capability to extract information and trends from large amounts of textual data (Tan, 1999; Feldman and Sanger, 2007; Fuller et al., 2011), giving an overview of the main issues discussed in the reports. It is based on techniques of data mining, machine learning and natural language processing applied to text. It strongly relies on computer programs and algorithms, thus it is supposed to overcome the problems associated with content analysis's reliability (Milne and Adler, 1999).

The international academic community trusts both techniques, which have been largely used with reference to the textual analysis of annual reports, earnings releases and other corporate documents (Li, 2010). Yet, these techniques appear to be different.

Looking at their origins and how the larger academic community has used them, it emerges that content analysis and text mining are usually seen as opposite methods, based on different theoretical assumptions and devoted to answering different types of research questions. On the contrary, a different stance suggests that past scholars and books on these research methods have overemphasised differences between the two techniques, without trying to understand possible overlapping areas. Moreover, thanks to IT improvements and cross contaminations between fields of knowledge, distinctions between these two techniques have been blurred. Thus, the following question is put forward: *Are content analysis and text mining truly irreconcilable research methods?*

In detail, this paper aims to shed lights on the nature of these two techniques, their similarities and differences, and how they can be used in a real research setting that requires students or researchers to identify variations in the amount of information provided in company reports. To increase clarity, the paper presents

the results obtained from the application of both content analysis and text mining to the same set of company reports. Results are then expressed in terms of number of words. Their examination allows us to discuss whether the two techniques may be considered as alternatives, i.e., two competing options, in the analysis of sustainability reporting, that a researcher may choose during the research design phase.

2. DIFFERENCES AND POSSIBLE SIMILARITIES BETWEEN CONTENT ANALYSIS AND TEXT MINING

After having searched in the Scopus database for academic contributions containing the term 'text mining' in the title, abstract or keywords, the first undeniable evidence of text mining is its sparse diffusion within the academic community due to its very recent development. Only 1,772 documents were found, with the first document dating back to 1998. On the contrary, the same search related to 'content analysis' returned 22,134 documents, with the first one dating back to 1939. Focusing on the frequency of usage, reference or citation of the two techniques in academic articles during the last decades (Fig. 1), it appears that content analysis is more common. However, if we calculate the average growth rate of documents using or citing text mining (50%), it indicates that this technique is experiencing a relevant uptake when compared to the average growth rate of content analysis (17%). Similar results are obtained if we calculate the global and annual growth rate.

The Scopus database also reveals the different origins of the two techniques (Tab. 1). On one hand, text mining has been developed within the field of Computer Science, where several articles still discuss technical improvements and software usage. Applications are found in soft disciplines like Social Sciences, Business and Management and Decision Sciences. On the other hand, content analysis is a technique born within the studies of Communication, Political Science and Sociology - all included in the wider field of sciences named Social Sciences (Yu et al., 2011). The main domains of application of this technique are Social Sciences and Business, Management and Accounting. In studies of Psychology, Medicine and Arts and Humanities, content analysis also appears, mainly to uncover concepts, values and behaviours of patients or service users.

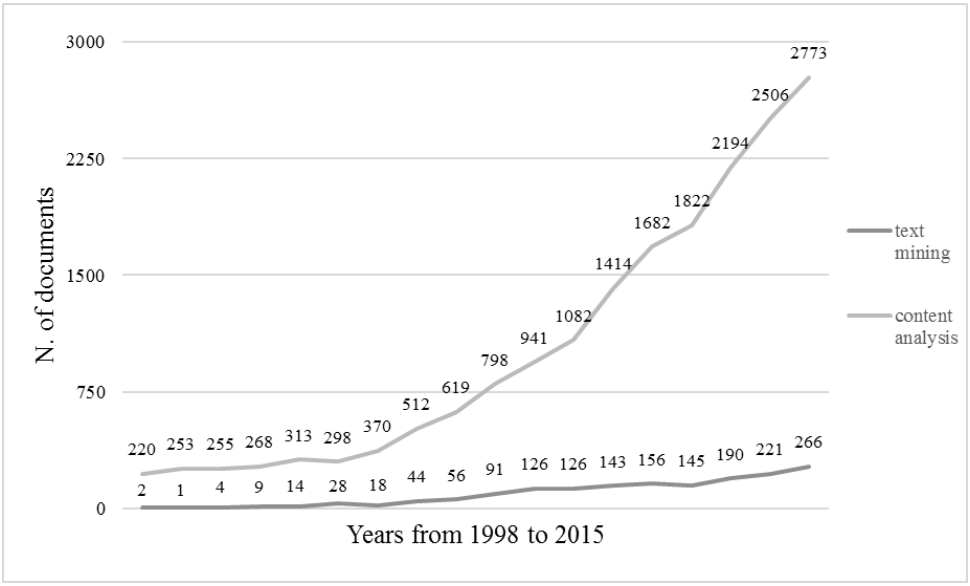

Figure 1. Diffusion of documents referring to the two techniques in the last decades

In these cases, notes from participant observation sessions and transcripts of interviews represent the basis for the text analysis.

Content Analysis		**Text Mining**	
70,4%	Social Sciences	51,0%	Computer Science
21,0%	Business, Management and	46,7%	Social Sciences
19,7%	Psychology	35,8%	Business, Management and
15,5%	Medicine	32,4%	Decision Sciences
10,5%	Arts and Humanities	12,6%	Engineering
8,1%	Computer Science	9,3%	Mathematics

5,7%	Economics, Econometrics and	7,9%	Arts and Humanities
3,8%	Nursing	6,3%	Psychology
3,2%	Engineering	4,2%	Economics, Econometrics and
3,0%	Decision Sciences	4,2%	Multidisciplinary
11,0%	Others	11,9%	Others

Table 1. Diffusion of the two techniques by subject area

The cited different origins are strictly linked to another important distinction between the two techniques, which is the respective assumption each holds about the nature of knowledge. In fact, a logical continuum exists between technical aspects and methodology.

Content analysis is diffused in disciplines where the qualitative research paradigm dominates. Coherently, this technique is traditionally used to uncover categories, patterns and relationships in the data by researchers that consider reality as a construct and adopt an inductive approach. From a hermeneutic perspective, content analysis is used to (re-)construct a 'reality' based on the interpretations of a text made by researchers. *"Coding forces the researcher to make judgments about the meanings of contiguous blocks"* (Ryan and Bernard, 2000, p. 785). Thus, the meaning of a text is negotiated among a community of interpreters (Patton, 2002). Moreover, the process of reading through the data and interpreting them continues throughout the project. The analyst may adjust the data collection process if additional concepts need to be investigated or new relationships need to be explored, following an iterative process. This and other distinctive aspects (Tab. 2) indicate that content analysis is a technique usually employed in interpretative studies and qualitative research.

On the contrary, text mining is rooted in the positivist research philosophy of "hard science" and seeks universal generalisations using measures or variables Rynes and Gephart (2004) and Lincoln and Guba (2000) for a description of

research traditions). Consequently, the role of the researcher and his/her subjectivity in analysing and making connections between sources of information should be limited. In qualitative content analysis, the researcher's knowledge guides the association of text segments to categories. The category system is developed right on the material but employs a theory-guided procedure, emerging from the synthesis of two contradictory methodological principles: openness and theory-guided investigation (Kohlbacher, 2006). Text mining neglects the active role of the researcher and aims to overcome the limits of content analysis linked to the variability of human categorisation of research topics/keywords. Text should be coded and analysed by the computer program and text is always transformed into numbers. In this case, the focus is not on meanings (behind words) but on how to better describe the world as it 'really' is and quantify phenomenon.

	Text mining	**Content analysis**
Research approach	quantitative	qualitative
Object	text	any type of qualitative data (text, image, video, etc.) but mainly used with text
Aim or intent of the research	explore/describe phenomenon AND/OR verify theories (hypothesis development)	explore/describe phenomenon AND/OR (eventually) develop meso-level theories
Unit of analysis	words/stems/combination of words	words/sentences/areas of pages
Instruments	with computer software	with or without computer software

Role of the researcher	passive (theory-free)	active
Type of process and literature usage	linear, based on defined steps: hypothesis developed by the literature which is then tested	iterative/circular; literature analysis can be missing or it can occur after data analysis
Outputs	tables and graphs	in text citations, tables and graphs

Table 2. Differences between the two techniques

Nevertheless, we cannot support such a strict distinction because techniques can be adjusted and used by researchers in different ways, following different methodological frameworks.

For example, content analysis can be classified within the quantitative research paradigm (Bryman and Bell, 2015) when used to develop measures and quantify phenomena. Instead of using an interpretative approach to content analysis, researchers may opt for a mechanistic approach (Beck et al., 2010), which focuses on the amount of words, sentences or other textual surrogates of meaning (Patten, 1991; Unermann, 2000; Patten and Crampton, 2003; Perrini, 2005) to measure the quantity of disclosure (Beretta and Bozzolan, 2004), to link disclosure with company performance (Clarkson, et al., 2008) and to test hypotheses on the relationship between social disclosure and other variables (Walden and Schwartz, 1997). Several recent accounting studies on social and environmental reporting have adopted this quantitative approach and use content analysis to generate quantitative measures of disclosure.

Because of the growing volume of corporate disclosure that limits manual approaches to text analysis (Indulska et al., 2012; Matthies and Coners, 2015) and the advent of more advanced computer programs for content analysis, the usage of this technique as a tool to quantify disclosure becomes more and more common. Operated with a quantitative approach in mind, content analysis appears to be similar to text mining. It becomes more positivistically oriented and shares the same assumption of text mining: the higher the frequency or occurrence of a

specific word or concept in a document, the greater the emphasis is that the author of the document or interviewee places on it. The text material is examined for the presence, frequency and centrality of concepts (also called conceptual analysis). These concepts are closely derived from the words used in the text because the approach used in coding is form oriented (focusing on the manifest meaning) and researchers do not search for latent structures of sense.

Moreover, increasingly popular computer programs are blurring the distinctions between quantitative and qualitative approaches to text analysis (Schutt, 2011). These software programs for content analysis include algorithms that help test hypotheses on relationships among concepts specified by the researcher as in theory-driven research studies. For example, the software may check for the co-occurrence of concepts by examining how these concepts are related to each other within the documents (this is called relational analysis).

The downside is that young researchers trying to decide which technique to adopt for their research project may get confused. Differences and mixture in terminology usage in journal articles add additional complexity. For example, some authors use a "*data mining-based content analysis tool*" (Indulska et al., 2012) which seems to include both techniques. Other authors (Li, 2010) use the term content analysis to describe a data collection technique based on a computer program, which automatically searches and classifies words based on a dictionary. This dictionary approach ignores any prior knowledge of the researcher and seems to be more similar to text mining than content analysis.

At the same time, we cannot state that text mining and qualitative research are epistemologically incompatible since several researchers use text mining as a viable qualitative research method (Camillo et al., 2005; Hong, 2009; Janasik et al., 2009; Yu et al., 2011). Text mining is similar to content analysis when both techniques are applied in line with grounded theory (Strauss and Corbin, 1990) to encourage open-mindedness and discourage preconceptions in researchers. For example, some text mining software programs can function without a pre-defined dictionary and identify concepts as a collection of words that co-occur with a certain frequency on the basis of a Baysian co-occurrence matrix (Indulska et al., 2012). In this case, text miners conduct the coding using automated algorithms,

which restrain the researcher from making any premature decisions in the research process (Yu et al., 2011). In addition, text mining can follow an interactive process like content analysis because the text miner can decide to add, delete, and revise the initial categories in an iterative fashion.

Similarities between the two techniques are even greater when the software for text mining includes natural language processing and the use of n-grams to extract content. This avoids the traps of focusing on single words or stems (stems are words lemmatised to reduce morphological variants, for example, by removing 's' and 'es' to convert plural into singular). When the software searches for combinations of more words (called n-grams), it is possible to capture key phrases and themes that provide context for entity sentiment and to identify specific aspects (e.g., "human right" becomes a single concept). Generally speaking, the lower the value of "n", the more general the phrase. Bi-gram (composed of two words) and tri-grams (three words) allow the software to better analyse complex texts as a human coder.

Figure 2 shows the aspects and modes of use that make the two techniques approach one another.

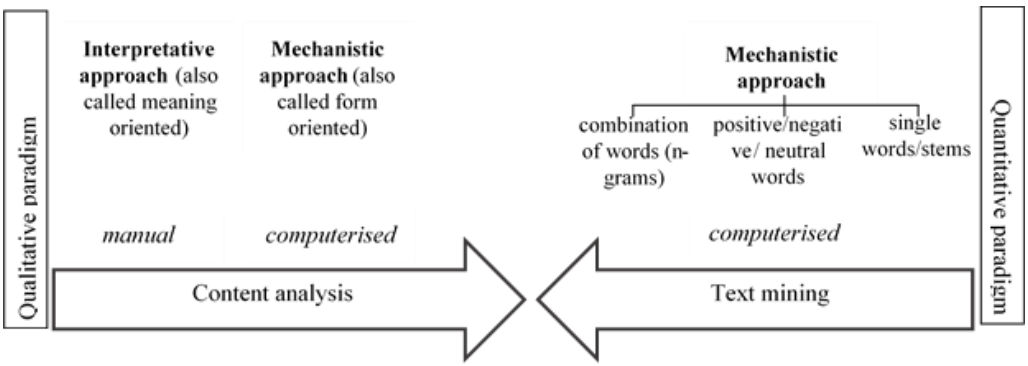

Figure 2. Variations in usage of content analysis and text mining

However, we should notice that there is no real uncontaminated view in any research involving text analysis. The research process is influenced by some preconceptions in both text mining and content analysis application. The text mining algorithms necessitate some conceptual structures about the proper classification method when we analyse two or more words. Text and data mining

software provide an automated approach to the coding of *corpi* but *"they do not replace the need for human interpretation of the coding"* (Indulska et al., 2012, p. 53). Researchers also make choices about the selection of stop words, stemming and dimensionality reduction, which may lead to variations in results. This is quite similar to what happens in content analysis, whose activities of coding and relationships building among concepts need some preliminary framework. Thus, there is no technique that can be labelled as purely objective.

Another aspect shared by both content analysis and text mining refers to the use of pre-defined coding schemes. For example, within content analysis, we may use a pre-existing framework made by different categories or concepts to capture the elements included in text. These categories may be sourced from previous studies or standards or relevant documents such as GRI guidelines. Studies aiming to quantify social and environmental disclosure often adopt this approach (Gray et al., 1995; Walden and Schwartz; 1997; Garzella and Fiorentino, 2013), which constrains the activity of researchers: they rely on a limited amount of categories into which content data must be coded. This procedure is quite different from free coding, which means that concepts are freely developed by researchers while reading text. Both options (pre-defined and emergent concepts) are possible, but the second one is more common in qualitative studies adopting an inductive approach when there are no theories that help to interpret a new phenomenon.

A similar choice is also available to researchers using text mining. For example, instead of directly using the software to analyse the text, researchers may adopt a pre-existing scheme of analysis, which consists in a list of words. In this case, researchers develop a vocabulary or dictionary, which is relevant for the study (Aureli et al., 2016). The vocabulary is drawn from previous studies or extracted from the text of regulations and standards. The vocabulary can even include a pre-defined list of positive and negative words, e.g., benefit is a word that the software counts as positive. Obviously, the vocabulary represents a limitation (as the use of pre-defined concepts), which may hinder the discovery of unexpected elements, but it also reduces the complexity of data collection and coding.

Nevertheless, content analysis seems to persist as the preferential data collection method for social and environmental issues reported. Researchers who adopt a

qualitative approach suggest that one of the key advantages of content analysis is to create countless classifications schemes and to allow the researcher to take context into account. For example, researchers can create categories related to the location of a topic within documents (e.g., in a letter to shareholders or in some annex) and categories to make the visualisation in tables and figures of relevant information explicit, which all represent additional sources of information about the relevance of themes disclosed. Moreover, researchers can code sentences as positive or negative regardless of the words used in the content; as interpreters of text, they can distinguish when the disclosure refers to positive or negative consequences related to company decisions and actions (Deegan and Rankin, 1996). In general, researchers can create categories to track the description of aspects and personal interpretations that are somehow related to the topics discussed (e.g., the inclusion of a comparison with what competitors do). These may help researchers answer different and/or more interesting questions such as which factors favour sustainability disclosure or what the circumstances related to company reporting are. However, doubts about reliability and validity emerge when these subjective classifications are used to quantify phenomena by counting words in the coded sentences.

3. METHODOLOGY

In order to understand whether the two techniques are more similar than one might suppose, this paper presents an experiment performed regarding the analysis of sustainability information reported by four different large corporations. The aim of the experiment is to investigate what happens to sustainability disclosure in the case of industrial disaster events that may generate a problem of social legitimacy and to check for any differences in how companies disclose sustainability before and after the event. In other terms, the question related to the experiment that both techniques should be able to address is: Does company disclosure on economic, social and environmental aspects increase or decrease after an industrial disaster?

The aim of the experiment implies a quantitative methodological framework. Both content analysis and text mining are applied to perform a trend analysis. Thus, their usage is intended to provide a description of what happens to sustainability

reports, without drawing any inferences from reasons for changes or related impacts.

The focus is on industrial disasters with global media coverage that have had serious repercussions for the companies involved in terms of reputation and those that have had a major impact on the academic world (also in terms of published studies dealing with the disaster). The companies under investigation are: *British Pretroleum* (BP), responsible for the marine oil spill disaster in the Gulf of Mexico on April 20, 2010; *Carnival Corporation*, holding of the Italian company responsible for the sinking of the Costa Concordia cruise ship in the Mediterranean Sea on January 13, 2012; *Lonmin*, a mining company located in South Africa, responsible for the killing of its workers during a strike on August 16, 2012; and *Tesoro*, one of the largest refineries in the United States that had to face the explosion of one plant causing deaths and the emission of harmful substances into the environment on April 2, 2010.

The two techniques have been applied to 20 documents in order to consider 5 years of disclosure for each company (Tab. 3). All of them, except for the Tesoro corporation, publish a sustainability report. Thus, for Tesoro, sustainability information has been searched for in annual reports. Considering the time society needs to become aware of the gravity of an event/disaster and make sustainability claims on the company (Summerhays and De Villies, 2012; Deegan and Islam, 2014), the empirical investigation assesses the reports regarding the year of the disaster (named year 0) and the following year (year 1). The three years before the disaster have also been considered to gain a clear picture of the disclosure trend.

Number of pages of each annual sustainability report				
Year	**BP**	**Carnival**	**Lonmin**	**Tesoro**
-3	44	46	103	115
-2	28	100	119	123

-1	40	86	81	100
0	50	66	32	126
1	54	104	197	140

Table 3. The documents analysed

The research steps required in the application of the two techniques are presented in Table 4. Driven by the necessity to identify and quantify disclosure related to sustainability aspects, existing and widespread key concepts have been used to define what refers to sustainability: those provided by GRI guidelines. GRI is the most widespread sustainability disclosure framework (KPMG, 2015; Thijssens et al., 2016), which underlines seven relevant aspects for corporate disclosure: Economic, Environmental, Labour Practices, Human Rights, Product Responsibility, Society and Strategy aspects. Each aspect is described in a specific section of the GRI document, which represents a relevant dimension of sustainability.

Recurrent words included in each GRI section were used to create the glossary employed to analyse company reports with the text mining software. With reference to content analysis, it is fundamental to discuss ideas, concepts and examples included in the GRI sections when more than one researcher or coder is involved. Researchers/coders should share their own reflections on GRI sections, using them as a point of reference or framework to code the text disclosed in company reports. Moreover, when more than two coders are involved, it is important to check for an intercoder agreement before starting the coding of text. Regardless of the software used, it is important to check whether the two coders categorise texts similarly. A sample document is usually coded independently by the two researchers. Then researchers compare the length of sentences selected and the category/concept associated to the sentence. If there are differences, these are discussed to improve the similarity in coding. Then the same exercise is repeated until a satisfying level of similarity is reached. In this experiment, coding means selecting text (a phrase or sentence, usually made by several words) which

refers to concepts held as important because they are related to the seven sustainability aspects identified by the GRI.

Researchers may choose among several different software for both content analysis and text mining. NVivo, Atlas.it, MAXQDA and Dedoose are commonly used for content analysis, while WordStat, Sas Text Analytics and Carrots2 are examples of text mining software. This choice is important, especially in reference to the text mining software. A single word can mean different things and the software should be able to detect these differences. For example, "books" in the phrase "he books tickets" is different from the same word in the phrase "he reads books" (Yu et al., 2011). In this experiment, preference has been given to Atlas.ti and the R text mining package, because they are largely known in academia, they are both free (although the free version of Atlas.ti has some limitations), there is a large amount of open access papers on how to use them and they are quite simple to use.

Results from text mining refer to the number of times the words in the sustainability glossary have occurred in each report. However, to reduce the risk of distortion due to the different lengths of the documents, the absolute frequencies (frequency of occurrence of each word) have been related to the total number of words in the document. Stopwords which generate confusion and do not add information - such as the, is, at, which, and, on, etc., - were excluded in the preparation of the relevant vocabulary. Results from content analysis can be displayed in terms of number of quotations (i.e., sentences picked by coders) and number of words included in the selected quotations. The software presents these results in tables called quotation count and word count, respectively. With respect to the two alternative output tables, it is more appropriate to focus on the amount of words related to sustainability aspects in order to compare content analysis (CA) results with those provided by text mining (TM) processing. Coherently, the calculation of the weight of sustainability-related words in relation to the total number of words included in the document was performed and stopwords have been excluded. Most of the software for content analysis entails the counting utility, allowing researchers to include or exclude a pre-defined list of stopwords.

	text mining	content analysis
PREPARATION	• Definition of the Framework: GRI guidelines (pdf file, version 3) • Identification of GRI sections (sustainability dimensions) and conversion into txt file of each section • Identification of key words for each section/dimension (stopwords removed) • Selection of the most recurrent key words and inclusion of synonyms (*Thesaurus*) • Conversion of key words and synonyms to stems which make the relevant vocabulary	• Definition of the Framework: GRI guidelines (pdf file, version 3) • Identification and discussion of GRI sections (sustainability dimensions) and identification of the categories/concepts included here that will be used to classify the text in reports • Definition of the common coding rules (i.e. sentence as a unit of analysis, no words, no tables and figures) • Reliability checks (e.g. intercorder agreement)
EMPIRICAL ANALYSIS	• Application of the vocabulary to sustainability reports with R tm package • Display results • Creation of tables	• Coding sustainability reports with Atlas.ti computer software • Display results • Creation of tables

Table 4. Research steps

4. DATA ANALYSIS

According to Table 5, both techniques suggest that Tesoro is the company with the lowest amount of sustainability words reported in its documents. This result is quite predictable because, in this experiment, Tesoro's text analysis has been performed on annual reports, which do not usually entail a lot of details on social and environmental aspects. Results from Carnival are also similar: this is the most active company in providing information on sustainability. On the contrary, the ranking of Lonmin and British Petrolum based on the average presence of sustainability-related words is different. Another common finding refers to the variability in the amount of words used throughout the years. Both techniques suggest that Tesoro is more stable in its disclosure (the historical series for this

company is particularly 'flat' and the standard deviation is low), while other companies vary the used quantity of sustainability terminology to a larger extent.

Presence of the glossary with TM					Presence of words coded with CA			
year	**BP**	**Carnival**	**Lonmin**	**Tesoro**	**BP**	**Carnival**	**Lonmin**	**Tesoro**
-3	13,94%	15,14%	15,36%	8,82%	34,49%	37,89%	12,87%	9,94%
-2	12,23%	16,21%	16,51%	9,03%	42,81%	45,87%	20,12%	10,64
-1	13,64%	16,83%	15,13%	9,41%	26,21%	29,17%	23,70%	8,14%
0	13,14%	15,28%	14,53%	9,42%	26,81%	19,98%	39,28%	8,31%
1	13,85%	16,66%	13,78%	9,40%	20,78%	23,45%	38,74%	6,89%
Mean	*13,36%*	*16,02%*	*15,06%*	*9,22%*	*30,22%*	*31,27%*	*26,94%*	*8,79%*
Std.	*0,71%*	*0,78%*	*1,01%*	*0,27%*	*8,57%*	*10,60%*	*11,69%*	*1,50%*

Table 5. Results Comparison

The more relevant differences emerge with reference to the disclosure trend of the companies analysed. For example, Figure 3 clearly indicates that the two techniques may describe a different story and lead to different conclusions on how companies react to an industrial disaster. In the case of British Petroleum, on one hand, content analysis suggests that the discourse on sustainability aspects increases the second year of observation (year -2), while it drops by almost half in the following years. A possible reaction to the disaster can be devised only in year 1 when the company reduces its disclosure on sustainability aspects from 26,81% to 20,78%. On the other hand, text mining describes an opposite trend. In year -2 the frequency of sustainability words decreases, while it grows in the following periods to return to previous levels. In addition, the disaster seems to have led to a different reaction: BP's disclosure increases in year 1 compared to the previous year (i.e., from 13,14% to 13,85%). Opposite results emerge with reference to the other companies too (Fig. 3). For example, according to text mining, Lonimin

reduces its reference to sustainability terms throughout the whole period, while content analysis suggests a growing trend for the disclosure of this company.

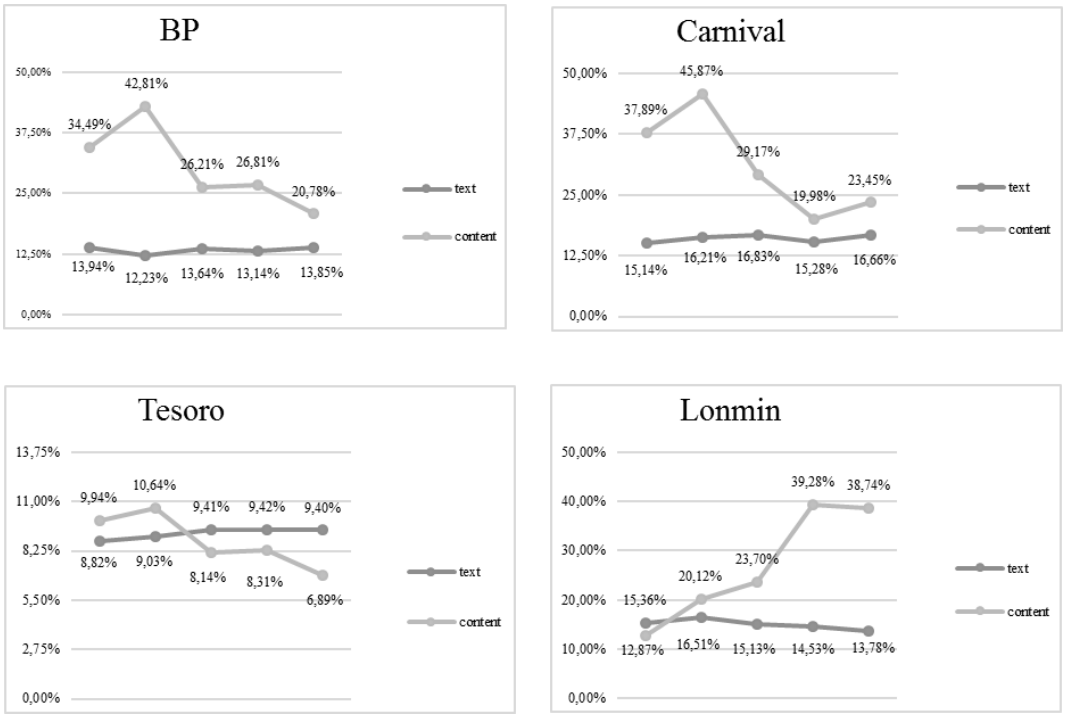

Figure 3. Trends in sustainability disclosure according to the different techniques

Thus, both techniques may provide an answer to our research question (what happens to the disclosure of sustainability aspects in the light of an industrial disaster event?), although leading to different conclusions.

Discrepancy is clearer in Table 6, which calculates annual variations in disclosure. Percentage of variations for the year of the disaster (first line - A) is obtained as followed:

$$\left(\text{Words freq}_{\text{year 0}} - \text{Words freq}_{\text{year -1}} \right) / \quad \text{Words freq}_{\text{year -1}} \quad = \ \%$$

Results indicate that content analysis and text mining describe a similar behaviour limited to Tesoro and Carnival. The first company decreases its disclosure in the year after the disaster (year 1), while the other one reduces its disclosure the year of the disaster (year 0), to then increase its discourse on sustainability in the

following period. On the contrary, variations in disclosure for BP and Lonmin present different signs.

	BP	**Carnival**	**Lonmin**	**Tesoro**
Text mining results				
A – Variation for the year of disaster (year 0)	-3,66%	-9,21%	-3,93%	0,01%
B - Variation for the year following the disaster (year 1)	5,42%	9,00%	-5,16%	-0,18%
C - Total variation - two year period	1,57%	-1,04%	-8,89%	-0,17%
Content analysis results				
A – Variation for the year of disaster (year 0)	2,31%	-31,49%	65,74%	2,12%
B - Variation for the year following the disaster (year 1)	-22,52%	17,36%	-1,37%	-17,14%
C - Total variation - two year period	-20,73%	-19,60%	63,47%	-15,38%

Table 6. Variations in company disclosure after the disaster

Looking at the four companies together, text mining suggests that there is an almost generalised reductive effect in disclosure produced by the negative event given that, in year 0, three out of four companies (BP, Carnival and Lonmin) recorded a percentage of the glossary, which is less than that obtained prior to the disastrous event. Content analysis suggests that the negative turning point in company disclosure is postponed to the year after the disaster (year 1), although, not all companies behave in the same manner.

5. DISCUSSIONS AND CONCLUSIONS

The performed experiment shows that both content analysis and text mining can be used as quantitative methods useful to solve the same research problem involving a trend analysis, which implies the transformation of company disclosure into quantitative variables that can be displayed with tables and analysed with statistical tools. However, the two techniques may lead to different findings. Thus, the two techniques cannot be used to crosscheck results.

Contrasting results in company behaviour may be attributed to the different way the techniques are used to identify references of sustainability aspects made in the text. Despite the usage of a common framework (GRI guidelines) in the experiment, text mining focuses on the use of single words, included in a vocabulary, that are considered specific or capable to detect sustainability issues, regardless of the context where these words are inserted, while content analysis focuses on sentences (sometimes longer, other times composed of a subject and a verb) describing a sustainability aspect. This is the reason for the lower percentages of frequencies detected by the text mining analysis compared to the count of word frequencies provided by content analysis.

At the same time, we should warn researchers that text mining's focus on words (stems) might be the cause of possible misinterpretations and abnormal situations. For example, text mining counts the term "climate" as many times as it appears in sustainability reports, without considering whether it is used to explain how the company handles climate changes or if it indicates the name of a new ERP software called "The Climate". On the contrary, the coder of content analysis selects sentences or phrases that describe a relevant issue – according to the researcher – no matter what the words used are. Thus, a quotation describing the company's will *"to reduce water consumption in the manufacturing of products as a mean to face the increasing lack of water springs*" is coded as a climate topic and included in the sustainability section of 'Environment' although the specific term "climate" is missing. Only a human coder can code different ad hoc articulations as the same concept.

Another explanation for the differences that have emerged in the experiment and that can be found in any other study, may be attributed to the fact that the same phrase can be counted two times in content analysis if its content deals with two different aspects at the same time. For example, the quotation describing a "*new policy that leads to both work safety improvements and reductions in the consumption of natural resources*" can be coded as 'Labour' and 'Environment' (two different sustainability sections) at the same time. This does not happen when performing text mining analysis, especially if the relevant glossary does not include any of the words used in the quotation. In the example provided, since the

word "work" belongs to the relevant glossary, it is included in the 'Labour' aspect, but no other terms, "consumption" nor "resource" increase the frequency of the 'Environmental' aspect, because none of these words belong to the vocabulary.

The cited differences do not depend on the commercial software used. They emerge because of the different approach of analysis embedded in each technique. Thus, results may lead to sustain that content analysis is a technique that better describes the meaning and relevance (expressed in terms of frequency) of certain topics because human coders, either coding the text manually or using a CA software to track concepts, are able to detect a concept no matter what words are used. It seems that text mining softwares takes on a simplified view of language that ignores the complexity of semiosis. Moreover, coders mimic the company's stakeholders who are the addressees of the communication. On the contrary, all software programs that rely on words (i.e., text mining software) entail some constraints due to possible limitations in the amount of words included (the dictionary embedded in the software may be old and not include technical terminology) and the construction of the relevant vocabulary based on the most recurrent words calculated using a cut-off rate.

In addition, text mining's focus on words may lead to claim that this technique merely describes a company's usage of specific terms and does not fully capture communications of a company's concerns about sustainability aspects. In other terms, one might argue that an increase in the sustainability glossary used may not be a good proxy for a company's behaviour and it may simply indicate a strategy of repeating disclosure that consists in inserting the same choice of words, several times, in different parts of company reports. Managers and/or preparers of the report usually repeat the same wording to produce amplification and to ensure that every reader will notice the company's words about sustainability aspects even when he/she focuses on a small portion of the report (Summerhays and De Villiers, 2012).

Consequently, content analysis should be preferred when the empirical research requires the identification of complex concepts that cannot be represented or synthesised by single words and when the content of reports, transcripts or other

text documents is highly unstructured, with irregularities and potential ambiguities.

In general, the preference for one technique over the other should be guided by the research questions, the several hypotheses that researchers aim to address and the type of data sources. For example, instead of analysing the amount of words communicated as in this experiment, researchers might be interested in understanding the extent and tone of disclosure, its presentation or the hidden meaning behind the words used. In this case, only a qualitative approach to content analysis allow researchers to identify patterns of impression management (a subtle form of influencing outsiders' perceptions of firm performance by manipulating the content and presentation of information in corporate documents) (Godfrey et al., 2003) and meanings formed by language.

From a methodological point of view, the experiment demonstrates that content analysis and text mining are not irreconcilable research methods, although the presence of possible overlaps between the two techniques is restricted to a specific research problem, which includes the transformation of text into numbers and measures.

With reference to companies' reaction to industrial disaster, the present experiment indicates that managers can implement different strategies. There is no mainstream or unique response. Increasing the amount of disclosure on economic, social and environmental aspects is a viable solution but the avoidance strategy can also be applied. They represent opposite reactions, both feasible.

Present comments are obviously confined to the four companies analysed. The small number of companies investigated represents the major limitation of this study. To formulate more robust conclusions it would be necessary to increase the company sample. Another limitation refers to the fact that the study does not address possible cross-country differences in the characteristics of corporate disclosure, assuming that GRI guidelines represent a standard that has set the tone and language for economic, social and environmental communication worldwide. Lastly, the empirical analysis does not consider other factors – such as management turnover – that might have an impact on company disclosure apart from industrial disaster. These additional factors that may cause a change in

disclosure are considered to be irrelevant since the final aim of this paper is to contrast results obtained from the two techniques and because sustainability reports are not strongly associated to a single top executive, given that they are the result of a common effort made by managers of different functions and departments, lawyers, public relations people and auditors (Li, 2010).

Future research should try to apply more advanced text mining instruments like natural language processing in order to verify if combinations of words – which are closer to human language – lead to results that are more similar to those obtained from content analysis. Thanks to continuous advancements in technology, text mining is becoming more and more sophisticated and may become a substitute for content analysis in the near future when researchers need to detect patterns in textual data and quantify the relevance of certain themes.

6. REFERENCES

ADAMS, C.A. (2002): "Internal organisational factors influencing corporate social and ethical reporting: beyond current theorising", *Accounting, Auditing & Accountability Journal*, vol. 15, n. 2: 223-250. DOI:10.1108/09513570210418905

ADAMS, C.A.; MCNICHOLAS, P. (2007): "Making a difference: Sustainability reporting, accountability and organisational change", *Accounting, Auditing and Accountability Journal*, vol. 20, n. 3: 382 – 402. DOI: 10.1108/09513570710748553

ADAMS, C.A.; LARRINAGA-GONZÁLEZ, C. (2007): "Engaging with organisations in pursuit of improved sustainability accounting and performance", *Accounting Auditing and Accountability Journal*, vol. 20 n. 3: 333–355. DOI: 10.1108/09513570710748535

ADAMS, C.A.; HILL, W.Y.; ROBERTS, C.B. (1998): "Corporate social reporting practices in Western Europe: legitimating corporate behaviour?", *The British Accounting Review*, vol. 30, n. 1: 1-21. DOI: 10.1006/bare.1997.0060

AERTS, W.; CORMIER, D. (2009): "Media legitimacy and corporate environmental communication", *Accounting, organizations and society,* vol. 34, n. 1: 1-27. DOI: 10.1016/j.aos.2008.02.005

AURELI, S.; MEDEI, R.; SUPINO, E.; TRAVAGLINI, C. (2016): "Sustainability Disclosure after a Crisis: A Text Mining Approach", *International Journal of Social Ecology and Sustainable Development,* vol. 7, n. 1: 35-49; DOI: 10.4018/IJSESD.2016010102.

BEBBINGTON, J.; LARRINAGA, C.; MONEVA, J.M. (2008); "Corporate social reporting and reputation risk management", *Accounting, Auditing & Accountability Journal,* vol. 21, n. 3: 337-361.

BECK, C.A.; CAMPBELL, D.; SHRIVES, P.J. (2010): "Content analysis in environmental reporting research: Enrichment and rehearsal of the method in a British-German context", *British Accounting Review,* vol. 42, n. 2: 207–222. DOI: 10.1016/j.bar.2010.05.002

BELLRINGER, A.; BALL, A.; CRAIG, R. (2011): "Reasons for sustainability reporting by New Zealand local governments", *Sustainability Accounting, Management and Policy Journal*, vol. 2, n. 1: 126 – 138. DOI: 10.1108/2040802 1111162155

BERETTA, S.; BOZZOLAN, S. (2004): "A framework for the analysis of firm risk communication", *The International Journal of Accounting*, vol. 39: 265– 288. DOI: 10.1016/j.intacc.2004.06.006

BONILLA-PRIEGO, M.J.; FONT, X.; PACHECO-OLIVARES, M.D.R. (2014): "Corporate sustainability reporting index and baseline data for the cruise industry", *Tourism Management*, vol. 44: 149-160. DOI: 10.1016/j.tourman. 2014.03.004

BONSÓN, E.; BEDNÁROVÁ, M. (2015): "CSR reporting practices of Eurozone companies", *Revista de Contabilidad – Spanish Accounting Review*, vol. 18, n. 2: 182-193. DOI: 10.1016/j.rcsar.2014.06.002

CAMILLO, F.; TOSI, M.; TRALDI, T. (2005): *Semiometric approach, qualitative research and text mining techniques for modeling the material culture of happiness*. Berlin, Germany: Springer.

CHO, C.H. (2009): "Legitimation strategies used in response to environmental disaster: A French case study of Total SA's Erika and AZF incidents", *European Accounting Review*, vol. 18, n. 1: 33–62. DOI: 10.1080/09638180802579616

CHO, C.H.; PATTEN, D.M. (2007): "The role of environmental disclosures as tools of legitimacy: A research note", *Accounting, Organizations and Society*, vol. 32, n. 7-8: 639–647. DOI: 10.1016/j.aos.2006.09.009

CLARKSON, P.M.; LI, Y.; RICHARDSON, G.D.; VASVARI, F.P. (2008): "Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis", *Accounting, Organizations and Society*, vol. 33, n. 4-5: 303-327. DOI: 10.1016/j.aos.2007.05.003

DE VILLIERS, C.; VAN STADEN, C.J. (2006): "Can less environmental disclosure have a legitimising effect? Evidence from Africa", *Accounting,*

Organizations and Society, vol. 31, n. 8: 763–781. DOI: 10.1016/j.aos.2006.03.001

DEEGAN, C. (2002): "Introduction: The legitimising effect of social and environmental disclosures-a theoretical foundation", *Accounting, Auditing & Accountability Journal*, vol. 15, n. 3: 282–311. DOI: 10.1108/09513570210435852

DEEGAN, C., RANKIN, M. (1996): "Do Australian corporate report environmental news objectively? An analysis of environmental disclosures by firms prosecuted successfully by the environmental protection authority", *Accounting, Auditing and Accountability Journal*, vol. 9, n. 2: 50–67. DOI: 10.1108/09513579610116358

DEEGAN, C.; ISLAM, M.A. (2014): "An exploration of NGO and media efforts to influence workplace practices and associated accountability within global supply chains", *The British Accounting Review*, vol. 46, n. 4: 397–415. DOI: 10.1080/0969160X.2015.1093779

ELKINGTON, J. (1998): *Cannibals with Forks: The Triple Bottom Line of 21st Century Business*. New Society Publishers, Stony Creek, CT.

FARNETI, F.; GUTHRIE, J. (2009): "Sustainability reporting by Australian public sector organisations: Why they report", *Accounting Forum*, vol. 33, n. 2: 89-98. DOI: 10.1016/j.accfor.2009.04.002

FELDMAN, R.; SANGER, J. (2007): *The Text Mining Handbook*. Cambridge University Press, New York.

FULLER, C.M.; BIROS, D.P.; DELEN, D. (2011): "An investigation of data and text mining methods for real world deception detection", *Expert Systems with Applications*, vol. 38, n. 7: 8392–8398. DOI: 10.1016/j.eswa.2011.01.032

GARZELLA, S.; FIORENTINO, R. (2013): "The Current Environmental Strategy Reporting Model: What can be Learned from Corporate Reports?", 121-137 in: MANCINI, D., VASASSEN E.H.D. AND DAMERI, R. (eds), *Accounting Information Systems for Decision Making*. Berlin Heidelberg, Springer.

GODFREY, J.; MATHER, P.; RAMSAY, A. (2003): "Earnings and impression management in financial reports: the case of CEO changes", *Abacus*, vol. 39, n. 1: 95–123. DOI: 10.1111/1467-6281.00122

GRAY, R., KOUHY, R.; LAVERS, S. (1995): "Corporate social and environmental reporting: a review of the literature and a longitudinal study of UK disclosure", *Accounting Auditing and Accountability Journal*, vol. 8, n. 2: 47-77. DOI: 10.1108/09513579510146996

GRAY, R.; MILNE, M.J. (2015): "It's not what you do, it's the way that you do it? Of method and madness", *Critical Perspectives on Accounting*, vol. 32: 51–66. DOI: 10.1016/j.cpa.2015.04.005

GUTHRIE, J.; ABEYSEKERA, I. (2006): "Using content analysis as a research method to inquire into social and environmental disclosure", *Journal of Human Resource Costing and Accounting*, vol. 10, n. 2: 114–126. DOI: 10.1108/14691930410533704

GUTHRIE, J.; FARNETI, F. (2008): "GRI sustainability reporting by Australian public sector organisations", *Public Money and Management*, vol. 28, n. 6: 361–366.

HOLCOMB, J.L.; UPCHURCH, R.; OKUMUS, F. (2007): "Corporate social responsibility: what are top hotel companies reporting?", *International Journal of Contemporary Hospitality Management*, vol. 19, n. 6: 461 – 475. DOI: 10.1108/09596110710775129

HONG, C.F. (2009): "Qualitative chance discovery – Extracting competitive advantages", *Information Sciences*, vol. 179: 1570-1583. DOI: 10.1016/j.ins.2008.11.041

INDULSKA, M.; HOVORKA, D.; RECKER (2012):" Quantitative approaches to content analysis: identifying conceptual drift across publication outlets", *European Journal of Information System*, vol. 21, n. 1: 49-69. DOI: 10.1057/ejis.2011.37

ISLAM, M.A.; DEEGAN, C. (2010): "Media pressures and corporate disclosure of social responsibility performance information: A study of two global clothing

and sports retail companies", *Accounting and Business Review*, vol. 40, n. 2: 131–148. DOI: 10.1080/00014788.2010.9663388

JANASIK, N.; HONKELA, T.; BRUUN, H. (2009): "Text mining in qualitative research: Application of an unsupervised learning method", *Organizational Research Methods*, vol. 12: 436-460. DOI: 10.1177/1094428108317202

KPMG (2013). *The KPMG Survey of Corporate Responsibility Reporting 2013*. KPMG International, Netherlands. https://assets.kpmg.com/content/dam/kpmg/pdf/2015/08/kpmg-survey-of-corporate-responsibility-reporting-2013.pdf

KPMG (2015). *The KPMG Survey of Corporate Responsibility Reporting 2015*. Haymarket Network, Netherlands. https://home.kpmg.com/content/dam/kpmg/pdf/2015/12/KPMG-survey-of-CR-reporting-2015.pdf

KOHLBACHER, F. (2006): "The Use of Qualitative Content Analysis in Case Study Research", *Forum Qualitative Sozialforschung / Forum Qualitative Social Research*, Vol. 7, n. 1.

KRIPPENDORFF, K. (2004): *Content Analysis: An Introduction to Its Methodology*. Sage Publications, Thousand Oaks, CA.

LI, F. (2010): "Textual analysis of corporate disclosures: A survey of the literature", *Journal of Accounting Literature*, vol. 29: 143-165.

LINCOLN Y.S.; GUBA E. (2000): "Paradigmatic controversies, contradictions, and emerging confluences", 163-188 in DENZIN N.K., LINCOLN Y.S. (eds), H*andbook of qualitative research*. Thousand Oaks, CA: Sage.

MATTHIES, B.; CONERS, A. (2015): "Computer-aided text analysis of corporate disclosures - Demonstration and evaluation of two approaches", *International Journal of Digital Accounting Research*, vol. 15: 69-98. DOI: 10.4192/1577-8517-v15_3

MICHELON, G. (2011): "Sustainability Disclosure and Reputation: A Comparative Study", *Corporate Reputation Review*, vol. 14, n. 2: 79-96. DOI: 10.1057/crr.2011.10

MILNE, M.J.; ADLER, R.W. (1999): "Exploring the reliability of social and environmental disclosures content analysis", *Accounting Auditing Accountability Journal*, vol. 12, n. 2: 237–256. DOI: 10.1108/09513579910270138

MODAPOTHALA, J.R.; ISSAC, B.; JAYAMANI, E. (2010): "Appraising the Corporate Sustainability Reports – Text Mining and Multi-Discriminatory Analysis", 489–494 in: Sobh, T.; Elleithy, K (eds) *Innovations in Computing Sciences and Software Engineering*, Springer, Dordrecht. DOI: 10.1007/978-90-481-9112-3_83

NEUMAN, W. (2003): *Social Research Methods: Qualitative and Quantitative Approaches.* Allyn & Bacon, Boston, MA.

O'DONOVAN, G. (2002): "Environmental disclosures in the annual report: Extending the applicability and predictive power of legitimacy theory", *Accounting, Auditing & Accountability Journal*, vol. 15, n 3: 344–371. DOI: 10.1108/09513570210435870

O'DWYER, B. (2002): "Managerial perceptions of corporate social disclosure: An Irish story", *Accounting, Auditing & Accountability Journal*, vol. 15, n. 3: 406–436. DOI: 10.1108/09513570210435898

PATTEN, D.M. (1991): "Exposure, legitimacy, and social disclosure", *Journal of Accounting and Public Policy*, vol. 10 297–308. DOI: 10.1016/0278-4254(91)90003-3

PATTEN, D.M.; CRAMPTON, W. (2003): "Legitimacy and the internet: an examination of corporate webpage environmental disclosure", 31–57 in Jaggi, B; Freedman, M. (ed), *Book Series Advances in Environmental Accounting and Management*, Vol. 2. Emerald Group Publishing Limited. DOI: 10.1016/S1479-3598(03)02002-8

PATTON, M.Q. (2002): *Qualitative Research & Evaluation Methods*, 3rd Edition. Sage Publications, Thousand Oaks, CA.

PERRINI, F. (2005): "Building a European portrait of corporate social responsibility reporting", *European Management Journal*, vol. 23, n. 6: 611–627. DOI: 10.1016/j.emj.2005.10.008.

ROCA, L.C.; SEARCY, C. (2012): "An analysis of indicators disclosed in corporate sustainability reports", *Journal of Cleaner Production*, vol. 20: 103-118. DOI: 10.1016/j.jclepro.2011.08.002

RYAN, G.W.; BERNARD, H.R (2000): "Data management and analysis methods", in Denzin, N. & Lincoln, Y.S. (Eds.), *Handbook of qualitative research* (pp.769-802). Thousand Oaks: Sage.

RYNES, S.; GEPHART JR., R.P. (2004): "Qualitative research and the Academy of Management Journal", *Academy of Management Journal*, vol. 47, n. 4: 454-462. DOI: 10.5465/AMJ.2004.14438580

SCHUTT, R.K. (2011): "Qualitative Data Analysis" (Chapter 10) in: *Investigating the Social World: The process and practice of research*, 7th Edition. Sage Publications, Thousand Oaks.

STRAUSS, A.; CORBIN, J.M. (1990): *Basics of qualitative research: Grounded theory procedures and techniques*. Sage Publications, Thousand Oaks.

SUCHMAN, M.C. (1995): "Managing legitimacy: Strategic and institutional approaches", *Academy of management review*, vol. 20, n. 3: 571-610. DOI: 10.5465/AMR.1995.9508080331

SUMMERHAYS, K.; DE VILLIERS, C.J. (2012): "Oil company annual report disclosure responses to the 2010 Gulf of Mexico oil spill", *Journal of the Asia-Pacific Centre for Environmental Accountability*, vol. 18, n. 2: 103–130.

TAN, A.H. (1999): "Text mining: The state of the art and the challenges", Paper presented at the Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases.

TE LIEW, W.; ADHITYA, A.; SRINIVASAN, R. (2014): "Sustainability trends in the process industries: A text mining-based analysis", *Computers in Industry*, vol. 65: 393–400. DOI: 10.1016/j.compind.2014.01.004

TEWARI, R.; DAVE, D. (2012): "Corporate social responsibility: Communication through sustainability reports by Indian and multinational companies", *Global Business Review*, vol. 13, n. 3: 393–405. DOI: 10.1177/097215091201300303

THIJSSENS, T; BOLLEN, L; HASSINK. H. (2016): "Managing sustainability reporting: many ways to publish exemplary reports", *Journal of Cleaner Production* vol. 136: 86-101.DOI: 10.1016/j.jclepro.2016.01.098

UNERMANN, J. (2000): "Methodological issues-reflections on quantification in corporate social reporting content analysis", *Accounting Auditing and Accountability Journal*, vol. 13, n. 5: 667–681. DOI: 10.1108/09513570010353756

WALDEN,W.D.; SCHWARTZ, B.N. (1997): "Environmental disclosures and public policy pressure", *Journal of accounting and public policy*, vol. 16, n. 2: 125 -154. DOI: 10.1016/S0278-4254(96)00015-4

WEBER, R.P. (1988): *Basic Content Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-049. Sage, Beverly Hills, CA.

YU, C.H.; JANNASCH-PENNELL, A.; DI GANGI, S. (2011): "Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability", *Qualitative Report*, vol. 16, n. 3: 730–744.