Building an Environmental Sustainability Dictionary for the IT Industry

Qi Deng
Carleton University
qi.deng3@carleton.ca

Michael Hine
Carleton University
mike_hine@carleton.ca

Shaobo Ji
Carleton University
shaobo.ji@carleton.ca

Sujit Sur
Carleton University
sujit.sur@carleton.ca

Abstract

Content analysis is a commonly utilized methodology in corporate sustainability research. However, because most corporate sustainability research using content analysis is based on human coding, the research capability and the scope of the research design has limitations. The relatively recent text mining technique addresses some of the limitations of manual content analysis but its usage is often dependent upon the development of a domain specific dictionary. This paper develops an environmental sustainability dictionary in the context of corporate sustainability reports for the IT industry. In support of building said dictionary, we develop a standardized dictionary building process model that can be applied across many domains.

Keywords: dictionary building process, environmental sustainability, text mining, IT industry

1. Introduction

Research on corporate sustainability (CS) reporting has a long history of using a manual content analysis (MCA) method based on human coding [1-3]. This choice of methodology is based on many reasons. First, MCA is a well-established research method and a set of research procedures that has been developed and validated to guide the research process. Second, MCA has been widely used as a way to make valid inferences from textual data, which happens to be the main content format of corporate sustainability disclosures. Third, MCA is an alternative method to examine issues, which would be time and resource intensive and too obtrusive if studied using other techniques, e.g., direct observations. For researching corporate sustainability reporting, content analysis of what companies have disclosed regarding their sustainability performances might be the most effective and appropriate methodology [see 4-12]. However, with the large volumes and high velocity of digitized textual materials on corporate sustainability reports increasingly being made available, MCA becomes constrained and less efficient as it is extremely time consuming and prone to human error. Increasingly, MCA faces criticism about coding reliability and potential coding errors caused by coder fatigue, misapplication of the coding schema, and potential disagreement between coders on particular attribute values [13-15].

Text mining (or automated content analysis, ACA) has the potential to address these problems. Text mining refers to the process of detecting patterns or knowledge from unstructured or semi-structured text and it has many advantages over MCA, such as enhanced reliability, elimination of manual coding errors, low cost, and the capability for analyzing large amounts of data in considerably short time period [4, 16-18].

In many cases, text mining is reliant on a thesaurus-like dictionary. A typical dictionary includes categories that contain words, word stems, and phrases. The frequencies of the words, stems, phrases, and thus categories, are counted and, based on these frequencies, the relative importance or changes over time of the central concepts in the texts can be determined. The dictionary allows researchers to systematically assess different aspects of the core concept they are interested in. In dictionary based text mining efforts, the quality of the results is largely dependent on the quality of the dictionary. Developing a dictionary is an iterative and time-consuming process which could last from months to years [17, 19]. However, once developed, a dictionary can be applied to any text mining projects related to the same domain and is very useful for document indexing and categorization as well as document retrieval [19]. There is no doubt that research on corporate sustainability reporting can benefit from the text mining method, especially in view of the increased digital availability of large volumes of CS related textual data.

While some previous corporate sustainability research has applied text mining method, most efforts have been at an introductory level and no related dictionary has been developed [see 4, 15, 20]. Considering its potential capability and current wide adoption in other research areas (e.g., Tourism, Agriculture, Political Science, Medical Science, and Psychology), the most possible reason for this under-utilization is the lack of a valid dictionary. Thus, to facilitate future proliferation of text mining in corporate sustainability research the necessary first step is to develop a useful and valid dictionary. While several papers using text mining touch upon the problem of building dictionary, in most, if not all, of them, the processes used to build the dictionary are not described clearly and are more or less subjective. To the best of

HICSS

our knowledge, no systematic dictionary building process has been documented in any manuscript.

This paper has two main objectives: 1) to develop a general dictionary building process model; 2) to actualize the aforementioned process in building a dictionary for detecting environmental sustainability topics for IT companies and demonstrate the initial dictionary's usage. The rest of the paper is organized as follows. Section 2 presents a dictionary building process model developed based on a review of previous related research. The method used to build the environmental sustainability dictionary and the result are described in Section 3. Section 4 presents a demonstration of using the dictionary to analyze the newspaper articles. Section 5 presents the discussion and conclusion.

2. Dictionary Building Process

To build a dictionary, one needs to identify the "right" words and/or phrases in the corpus and assign them into different categories that represent concepts that the researcher is interested in. For example, to build an emotion dictionary which can be used to analyze online product comments, researchers probably identify the words "satisfy", "good", and "useful" as representative of positive emotion and the words "terrible", "angry", "useless" as that of negative emotion.

Previous literature on dictionary building has two streams: automatic dictionary building and semi-automatic dictionary building. Rooted in information extraction research, automatic dictionary building usually involves extracting key words and/or phrases automatically based on learning algorithms and evaluating the resulting dictionary by experiments or comparing it with existing dictionaries [see 21-26]. In semi-automatic dictionary building researchers make their own dictionary inclusion judgements on words and/or phrases with the assistance of text analysis software.

In this paper, we develop a process model to support dictionary building consistent with the existing semi-automatic dictionary building research. A preliminary search of literature on text mining and addressing dictionary building resulted in 15 papers. None of these papers adopted a standardized dictionary building process as is common in automatic dictionary building research. Following an inductive approach, where possible, we: analyzed the descriptions of dictionary building processes (or lack thereof) in these papers, summarized the steps adopted (see Table 1), and subsequently derived a general dictionary building process.

Table 1. Summary of dictionary building process

Citation	Dictionary	Corpus Creation	Pre-processing	Entry Identification & Categorization	Extension & Simplification			Validation
					Synonyms & Antonyms	Stemming	Weighting	
[27]	Online image and video subject	X		X				
[28]	Job description	X		X				
[29]	Tone in financial text	X		X				
[30]	Corporate philanthropy	X		X				
[31]	External validation, shareholder alignment, market performance and accounting performance	X		X				X
[32]	Rational and normative words	X		X				X
[33]	Precautionary principle	X		X				X
[34]	Forest value	X		X				X
[35]	Auditing research topics	X	X	X				
[36]	Danish Adverse Drug Events	X	X	X				X
[37]	Competency-related terms in business intelligence and big data job ads.	X	X	X			X	
[38]	Privacy related issues	X	X	X	X			X
[39]	Privacy related issues	X	X	X	X			X
[40]	Policy agendas	X			X	X		X
[18]	Public leadership image	X				X		X

We name the resulting documentation the "semi-automatic dictionary building process (S-DBP)". The S-DBP includes five steps, namely, corpus creation, pre-processing, word and phrase (entries) identification and categorization, extension and simplification, and validation. While iteration within the steps is common we will discuss the steps in linear fashion.

Step 1. Corpus creation. The corpus is the source documents from which the dictionary is developed. It usually consists of multiple documents which include rich textual contents related to the topic of the dictionary. Creating a corpus involves selecting the right textual sources for future processing. Since the dictionary is derived from the corpus, its quality is directly dependent on the documents in the corpus.

Previous studies have not generally addressed the assessment of corpus. Three features of the corpus could be considered to decide whether the corpus is "adequate". First, the corpus should be relevant. It should include the contents which are consistent with the theme of the dictionary to be built. Second, the corpus should be appropriate. Since the subsequent steps are mainly based on the analysis of text, the original corpus should include mainly textual contents, instead of numeric or pictorial contents. Third, the corpus should be complete. For example, in order to build a dictionary of forest values, Bengston and Xu [34] created a corpus which includes articles by forest economists, traditional foresters, forest ecologists, landscape architects, aestheticians, environmental philosophers, environmental psychologists, Native Americans, and so on. To be complete does not mean that the corpus should include every related document, rather its should ensure that the richness and completeness of the corpus should be adequate to support the dictionary building. The criterion of "completeness" is especially important for the process of a building dictionary with pre-specified categories. If the corpus does not cover all pre-specified categories, neither will the dictionary.

Step 2. Pre-processing. The aim of this step is to prepare the corpus for further analysis using data cleaning techniques including: stop words removal [see 37], unnecessary information removal [see 35-36], reducing phrases to single words [see 38-39], spelling correction, and so on. Usually the pre-processing is conducted with the help of text analysis or text mining software. Currently, there are many computer-aided text analysis (CATA) software can assist with the pre-processing step, such as *WordStat* and *RapidMiner* among others. Whether to conduct this step and which techniques to be used are decisions which are made by researchers based on the requirement of the dictionary. Of the 15 identified papers, 5 include this step and 10 do not conduct this step.

Step 3. Entry identification and categorization. Usually, a dictionary includes three basic elements: the entries (words, word stems and phrases), the categories, and the association between the entries and the categories. Categories, according to Weber [41, p. 140] are "*a group of words [and phrases] with similar meaning and/or connotations*". In this step, researchers, who are familiar with the theme of the dictionary, examine each entry in the list developed in the second step and decide whether the entry should be retained and into which category the entry should be assigned. Entry identification and categorization are typically carried out by researchers with assistance of text analysis software. Many projects do not have pre-specified categories and are more exploratory in nature. In these situations, dictionary categories are derived from the content of the corpus itself. Typically, this is done with the aid of a 'topic extraction' feature within text mining software that aids in uncovering thematic structure of the processed text. Topic extraction is usually implemented using latent semantic analysis or latent dirichlet allocation.

Researchers often determine cut-off criteria and exclude entries from the dictionary that do not meet the criteria. Popular cut-off criteria include term frequency, and frequency of the documents in which one entry occurs. For example, "terms occurring in less than 1% of the documents" was used in Lesage & Wechtler [35] and Debortoli, Müller & vom Brocke [37] as cut-off criterion, while "terms occurring more than 30 times" was used in Abrahamson & Eisenman [32] as cut-off criterion. TF*IDF is another popular cut-off criterion. TF refers to term frequency and IDF refers to inverse document frequency. Although TF*IDF has not been used in the papers we reviewed, it is a standard way of culling words up front. The usage of this metric is based on the assumption that the more frequent a term occurs in a document, the more representative it is of the document's content yet, the more documents in which the term occurs, the less important the term is in distinguishing different documents' content from each other. So, if the purpose of the research is to distinguish between documents, as it is in classification tasks, TF*IDF is extremely important.

As our review indicates, the cut-off criterion is usually an arbitrary decision made by researchers based on the scope of the corpus or a decision to follow established criteria levels from previous studies. In most of the studies we reviewed, the entry identification and categorization are conducted by single researcher. However, it can be performed by multiple researchers as well. In the multi-coder case, the concept of inter-coder reliability is introduced as an assessment of the word categorization [see 32]. The result of this step is an

initial dictionary which could be further modified or directly applied to analyze additional text documents.

Step 4. Extension and simplification. Many techniques can execute this task, but generally speaking, the most common ones are synonym and antonym extension, stemming, lemmatization and weighting. The synonyms and antonyms extension means to add synonyms (and antonyms) for the initial words to the dictionary. Because of the various wording preference, different terms might be used by different authors to express the same meaning. To extend the dictionary by including synonyms (and antonyms) can, in some degree, increase the generalizability of the dictionary. The entries in the dictionary are not necessarily whole words or phrases, but are often reduced by stemming or lemmatizing. Stemming is a more rudimentary approach where words are simply truncated. For example, the word "having" maybe stemmed to "hav*". Alternatively, lemmatizing aims to retain the morphology of the word and would thus reduce "having" to "have". The choice of approach is project dependent. Stemmers are faster and simpler but lemmatization is more accurate. In this way, the dictionary can be simplified without costing the accuracy and effectiveness. Weighting means to weight terms based on their occurrence in and across documents. It is usually performed by applying the commonly used TF*IDF (Term Frequency-Inverse Document Frequency) weighting scheme. Compared with synonyms and antonyms extension, stemming, and lemmatization weighting is less commonly used. But, in some special cases this technique can promote the occurrence of rare terms and discounts the occurrence of more common terms [37, 42].

Step 5. Validation. The fourth step results in an extended and simplified dictionary that should be validated before being widely applied. Of the 15 studies reviewed, 9 report some form of validation of the dictionary. As the review shows, the most common validation method is to examine the key-word-in-context (KWIC), following by to compare-with-human-coding (CWHC), demonstration, and expert validation. Since the same entry might have different meaning in different context, it is necessary to have a look at the actual usage of the entry in the corpus to determine whether the entry is the accurate indicator of the concept the researcher perceives it to indicate. Another validation method is to compare the automated coding results with human coding results. The similarity between the automated coding results and human coding results are the primary indicator of the validity of the dictionary. Researchers also can validate the dictionary by demonstration (to actually apply the dictionary) or by expert validation (to have an expert on the theme of the dictionary to have a review of the dictionary).

One item of note is that the S-DBP aims to provide instructional guidelines, rather than impose requirements, for researchers interested in domain-specific dictionary building. Although we illustrate the dictionary building process as a sequential step-by-step process, in reality dictionary building is an iterative process where steps are often revisited. For example, if the quality and quantity of the entries identified in step 3 are below one's expectation, one might need to re-think about the corpus creation. After validation, one might need to re-think the whole process to see if there are any improvements one can do to make the dictionary better one. To build a comprehensive dictionary is a long-term activity which could last from months to years [19, 40, 43]. However, not every dictionary is necessarily comprehensive. The scope of the dictionary is decided based on the purpose of the research. The dictionary can be used confidently as long as it is comprehensive enough to support its purpose. In next section, we describe the process of building an environmental sustainability dictionary for IT companies following the S-DBP approach.

3. Environmental Sustainability Dictionary

We follow the S-DBP described above to build a dictionary for environmental sustainability of IT companies. With the rise of the concept of "Green IT", the IT industry has paid increasing attention to environmental sustainability. We use *WordStat* from *Provalis Research* to support the dictionary building process. *WordStat* has been used extensively in text analysis related research.

Step 1: Corpus creation. Corporate sustainability reports of IT companies from the *2015 Fortune 500* were collected and used to create the corpus for dictionary building for three reasons. First, corporate sustainability reports usually include economic, social, and environmental sustainability performance content; it is thus related. Second, despite the presence of some numerical data, most of the contents in the corporate sustainability report are textual data, and therefore appropriate. Third, the corporate sustainability report is one of the most important artefacts to communicate a company's sustainability performance to its stakeholders. Therefore, it generally includes every aspect of the company's sustainability performance and can be considered complete. Of the 49 IT companies included in the 2015 Fortune 500, 28 issued annual corporate sustainability reports, 10 issued online sustainability disclosures, and 11 did not disclose corporate sustainability information. To improve the corpus' relatedness, we only collect the environmental section from the CS reports and online disclosures from

2015. This results in 751 pages (reduced from 2,119 pages) of CS report contents and 53 pages of online disclosure contents. In total, the initial corpus consists of 38 documents (reports or online disclosures), which include 804 pages of environmental sustainability related contents.

Step 2: Pre-processing. After importing the initial corpus into *WordStat*, we conducted two steps of pre-processing; spelling check and stop word (e.g., "a", "and", "or", etc.) removal. Although corporate sustainability reports and online disclosures are official publications and, usually, they do not include spelling mistakes, it is still necessary to conduct a spelling check before further analysis because the format of the textual data might change during the data importing step. For example, the original phrase, "*environmental sustainability*", might become "*environnmentalsustainability*" after being imported. Since these format changes influence the frequency analysis later, it is necessary to deal with them before conducting next step. The spelling check can be conducted with the help of built-in functions of *WordStat*.

WordStat has a built-in stop word dictionary which includes common stop words and can be refined by researchers according to the research objective. Enabling the stop word removal function will automatically exclude the stop words from the subsequent text analysis. We used the default stop words dictionary because it does not include sustainability-related words, thus, will not impact the text analysis later.

Step 3: Entry Identification & Categorization. In this paper, we adopted an iterative process to identify and categorize the environmental sustainability-related entries. The 38 documents were randomly divided into a training set and a testing set, with each including 19 documents. We then developed an initial dictionary from the training set. We then refined the initial dictionary by applying it in the testing set to see whether there are qualified entries in leftover entry set. The testing set was randomly divided into four subsets (5 documents for three subsets and 4 for one subset) and the initial dictionary was refined through four rounds. Both the initial development and the later refinement followed similar entry identification and categorization process as described below.

Entry categorization. We adapted the environmental sustainability categories of the GRI G4 reporting framework to support the entry categorization. This approach is consistent with many corporate sustainability studies [see 3, 5, 7, 15]. The GRI G4 environmental sustainability framework covers twelve related aspects including: materials, energy, emissions, water, biodiversity, effluents & waste, products & services, compliance, transport, supplier environmental assessment, environmental grievance mechanisms, and overall. We remove "overall" from our categorization framework because it is fully overlapped with other categories. Therefore, we pre-specified eleven categories.

Entry identification. In the initial development stage, and after pre-processing, the 19 documents in training set contained 7,487 words. After applying the cut-off criterion of "occurring in no less than 2 documents", 3,865 words are retained. After applying the cut-off criterion of "occurring no less than 5 times with max words of 4", 915 phrases were generated. The first author then manually reviewed the 4,780 entries (both words and phrases) and identified environmental sustainability-related entries which represented the eleven categories of the coding schema described above. Each identified entry was categorized based on the examination of keywords in context (KWIC). The initial attempt resulted in a dictionary containing 261 entries. We then applied the dictionary in the testing subsets and examined the leftover words following the same cut-off criteria to see whether there were additional qualified entries. After four rounds of refinement, the dictionary included 287 entries.

Step 4: Extension & Simplification. For the words in the initial dictionary, we examined their synonyms, which also occur in the documents, to see whether they should be included in the dictionary. Similar to the initial coding, this step was also guided by the coding schema and with the help of KWIC. This step generated 15 new words. We did not conduct stemming or lemmatization here because we found that, sometimes, the different tenses of one word had different meanings. Finally, since this was the first step to build an environmental sustainability dictionary, we did not weight the entries either.

Step 5: Validation. We conducted two rounds of validation of the dictionary. In the first round, we designed a task of re-coding the previously identified entries into the dictionary categories. A PhD student, who was familiar with corporate sustainability concepts, was hired to conduct this task. The task included two rounds. In the first round, the student was asked to re-categorize the entries in the dictionary into the eleven categories based on our coding schema without the assistance of the KWIC capability. In the second round, the student was asked to perform the task with the help of the KWIC. In both rounds, the student did not know the original coding results of the entries. The reliability between original coding and additional coding is shown in table 2 below.

Table 2. Inter-Coder reliability of the dictionary validation

No.	Category	Number of Entries	Reliability	
			Round 1	Round 2
1	BIODIVERSITY	13	54%	62%
2	COMPLIANCE	23	100%	100%
3	EFFLUENTS & WASTE	45	53%	80%
4	EMISSIONS	38	82%	100%
5	ENERGY	73	75%	93%
6	ENVIRONMENTAL GRIEVANCE MECHANISMS	3	67%	67%
7	MATERIALS	27	48%	70%
8	PRODUCTS & SERVICES	24	63%	58%
9	SUPPLIER ENVIRONMENTAL ASSESSMENT	16	94%	88%
10	TRANSPORT	24	79%	79%
11	WATER	16	94%	100%
	All Entries	**302**	**72%**	**85%**

Scale of the inter-coder reliability: 0.21-0.40 (Fair); 0.41-0.60 (Moderate); 0.61-0.80 (Substantial); 0.81-1.00 (Almost Perfect) [44-45].

As shown in Table 2, the interrater reliability improved from 'substantial' (72%) to 'almost perfect' (85%) with the help of KWIC. The first author re-examined every entry coded differently from the second coder and discussed the entry context with the second coder. The dictionary was then refined based on the discussion. The final dictionary included 302 words and phrases, a portion of which are shown in Table 3. One thing to notice in this dictionary is that the entries are not equally distributed in different categories. The variety of the distribution reflects that the IT companies pay different attention to different environmental sustainability aspects. For example, it is clearly shown in Table 3 that IT companies have paid more attention to Energy, Emission, and Effluent & Waste than Biodiversity, Water, and Environmental Grievance Mechanisms. The demonstration of the generated dictionary is presented in the next section.

Table 3. Dictionary of environmental sustainability for IT industry (sample)

No.	Category	Entries
1	BIODIVERSITY	biodiversity; conservation; plants; tree; wildlife; …
2	COMPLIANCE	compliance; compliant; law; regulation; …
3	EFFLUENTS & WASTE	nonhazardous; composted; disposal; electronic waste; ewaste; landfill; product end of life; recycling; waste; remanufacturing; reuse; …
4	EMISSIONS	carbon offset; greenhouse; air emissions; air pollution; carbon; carbon dioxide; carbon neutral; dioxide; emission; footprint; …
5	ENERGY	air conditioning; biogas; cells; clean energy; cooling; electricity; energy; energy star; fuel; gas; gasoline; grid; heating; hydro; kilowatt; lamps; led; lighting; power; renewable energy; solar; wind; wind farm; …
…	… …	… … … …

4. Demonstration

The purpose of the demonstration is to show how the resulting dictionary can be used in an analysis of environmental sustainability for technology companies. Because of the small amount of data being analyzed and given the nascent stages of dictionary development we are cautious about drawing any conclusions from the results reported below. At this stage, we consider the demonstration as a "proof of concept" only.

For the demonstration, we collected environmental sustainability related newspaper articles from *LexisNexis*. To limit the scope for ease of demonstration, we only search related articles published in *New York*

Times from 2001 to 2015, which covers the 15 years during which the corporate sustainability achieved a rapid awareness worldwide. The method we used to search the articles is as follows:

HLEAD (Corporate Name, e.g., *Apple*, *Microsoft*, etc.) AND BODY (social responsibility) OR BODY (corporate responsibility) OR BODY (corporate citizenship) OR BODY (sustainability) OR BODY (environmental)

In total, the search results in 698 articles. Under some corporate names (i.e., Apple, Amazon), the search tends to result in more unrelated articles. The reason is that these searches result in some articles that are actually about apple, the fruit, and Amazon, the forest. We thus reviewed the first paragraph of each article to make sure that we only include sustainability related articles. This resulted in 449 articles. An import template was designed and the articles were brought into *QDAMiner / WordStat* for future analysis.

Using *WordStat* we detected all the words/phrases from the dictionary in the articles and generated a contingency table showing the percentage of words in each of the dictionary categories across year of publication. This data can then form the basis of analysis that adds insight into how the different topics (represented by categories) of environmental sustainability ebb and flow across time as reported by a media source. Because the outcome of the application of text mining is often a contingency table, it is typical to report results using correspondence analysis (CA). CA is a method that allows the graphical representation of contingency table data in low dimensional space [46]. CA has been successfully used in a variety of domains including marketing [47], tourism management [48-50], teaching and learning [51] among others.

While there are several types of CA maps available, Greenacre states that "the symmetric map is the best default map to use" (46: 267). The symmetric map typically provides a 'nicer-looking' representation than the asymmetric approach which often compresses the primary coordinates of the row profiles towards the centre of the map to allow the display of the extreme vertices of the column profiles (essentially creating a map that is more difficult to visualize than a symmetric map). The CA map of the contents of New York Times articles as detected by the sustainability dictionary is shown in Figure 1 below.
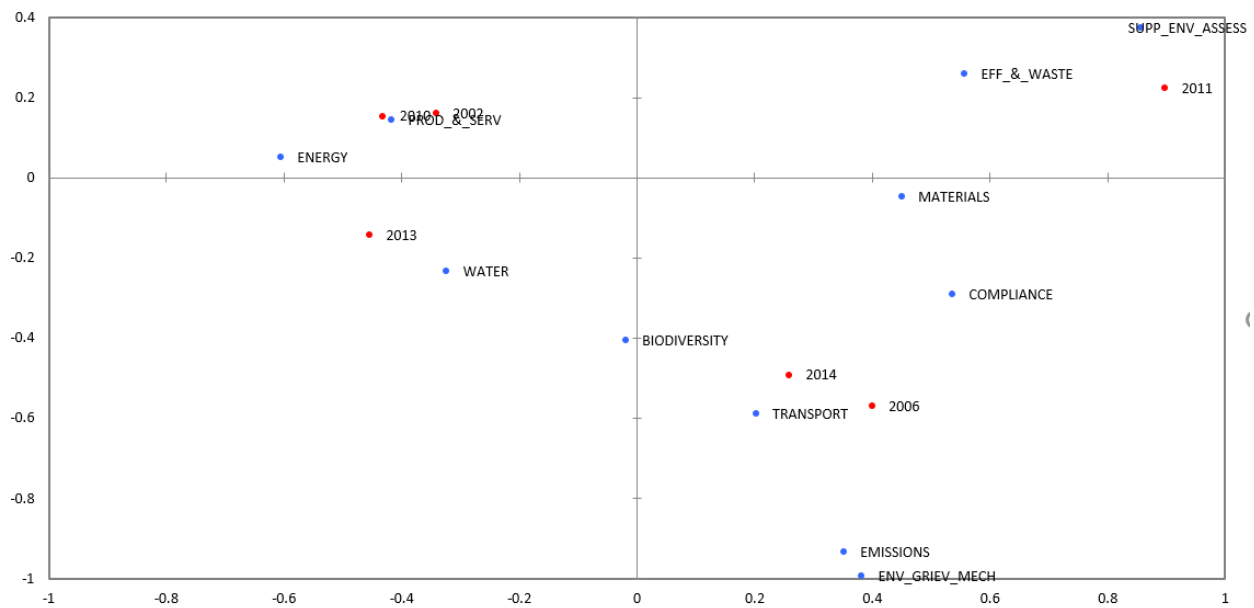

Figure 1. CA Map of environmental sustainability topics from the media across time

The point at which the axes cross represents the average yearly profile of environmental sustainability topics. Note that some years are not represented in the map as there was insufficient data to detect an appropriate amount of relevant words/phrases. If we look primarily at the horizontal axis, which in CA explains more of the variance than the vertical axis, we see that the yearly profiles are the most different between {2002; 2010; 2013} and 2011as the horizontal distance between these years is the greatest. The {2002; 2010; 2013} profiles are fairly similar but distinguished by proportionally more entries in water in 2013 and proportionally more entries on energy and 'products and services' in 2002 and 2010. The profiles of 2014 and 2006 articles are similar and proportionally contain more content related to transportation than do other years' articles. Finally, the 2011 profile is the most unique of the reported years with proportionally more entries dealing with 'effluent and waste' and 'supplier environmental assessment'.

5. Discussion and Conclusion

Corporate sustainability research can benefit from adopting a text mining approach. To promote and maximize the benefit, building a dictionary is a necessary first step. The aim of this paper is to take the first step. This paper has two major contributions. First, although research on dictionary building already exists, as far as we know, none of them follow or propose a standardized dictionary building process. Based on previous automated content analysis studies, we propose a semi-automatic dictionary building process, which includes five steps, namely, corpus creation, pre-processing, words identification and tagging, extension and simplification, and validation. Notably, the dictionary building is an iterative process which could last from months to years. Second, we have built an initial environmental sustainability dictionary for the IT industry. Although this dictionary is only an initial version and still need further modifications, we do believe that the development of such dictionary will promote the adoption of text mining method in corporate sustainability area and, in turn, facilitate the research in this area.

This paper is not without limitations. First, the corpus created for dictionary building is limited. We only included the most recent corporate sustainability reports (and online disclosures) in the corpus. Although, logically, the CS reports should cover every aspects of the companies' environmental sustainability activities, the dictionary building should probably incorporate data from different sources to ensure completeness. Despite of the data being sourced from company reports, future research could also incorporate data from mainstream media, non-profit organizations, government, among others. Second, during the dictionary building process, we made some arbitrary decisions. For example, we use the cut-off criterion of "occurring no less than 2 documents" without evaluating the impacts of the criterion on our results. As far as we know, previous research has not addressed the impacts of such cut-off criteria on the dictionary building results. However, for dictionary building research, such evaluation is significant. Future research could investigate that area. Third, due to the limitation of time and scope, we only included one extra coder to validate the dictionary. The increase of inter-coder reliability is not without an experience threat. Future research should include multiple coders and multiple trials to validate the dictionary. After development, the dictionary needs more robust validation. Since the quality of the text mining research is limited by the quality of the dictionary used, it is necessary and important to make sure the dictionary is adequate. To our knowledge, there is limited research addressing what might constitute an adequate dictionary [41]. We call for future studies to investigate this issue further.

In conclusion, the objective of the proposed S-DBP is to provide researchers interested in dictionary building with a general guideline to follow. Our hope is that the S-DBP could provide a basic model for future dictionary building. The second contribution of this study is the development of a dictionary for studying environmental sustainability in the IT industry. To our knowledge, it is the first dictionary developed for the corporate sustainability field.

6. References

[1] J.A. Arevalo, "Critical reflective organizations: An empirical observation of global active citizenship and green politics", Journal of Business Ethics, 96(2), 2010, pp. 299-316.

[2] M. Asif, C. Searcy, P.D. Santos, and D. Kensah, "A review of Dutch corporate sustainable development reports", Corporate Social Responsibility and Environmental Management, 20(6), 2013, pp. 321-339.

[3] Delai, and S. Takahashi, "Corporate sustainability in emerging markets: Insights from the practices reported by the Brazilian retailers", Journal of Cleaner Production, 47, 2013, pp. 211-221.

[4] R. Barkemeyer, L. Preuss, and L. Lee, "On the effectiveness of private transnational governance regimes – evaluating corporate sustainability reporting according to the global reporting

initiative", Journal of World Business, 50(2), 2015, pp. 312-325.

[5] M.J. Bonilla-Priego, X. Font, and M. del Rosario Pacheco-Olivares, "Corporate sustainability reporting index and baseline data for the cruise industry", Tourism Management, 44, 2014, pp. 149-160.

[6] J.M. Comas Martí, and R.W. Seifert, "Assessing the comprehensiveness of supply chain environmental strategies", Business Strategy and the Environment, 22(5), 2013, pp. 339-356.

[7] D. de Grosbois, "Corporate social responsibility reporting in the cruise tourism industry: A performance evaluation using a new institutional theory based model", Journal of Sustainable Tourism, 2015, pp. 1-25.

[8] C. Deegan, M. Rankin, and J. Tobin, "An examination of the corporate social and environmental disclosures of BHP from 1983-1997: A test of legitimacy theory", Accounting, Auditing & Accountability Journal, 15(3), 2002, pp. 312-343.

[9] A. Fonseca, "How credible are mining corporations' sustainability reports? A critical analysis of external assurance under the requirements of the international council on mining and metals", Corporate Social Responsibility and Environmental Management, 17(6), 2010, pp. 355-370.

[10] S.K. Fuisz-Kehrbach, "A three-dimensional framework to explore corporate sustainability activities in the mining industry: Current status and challenges ahead", Resources Policy, 46, 2015, pp. 101-115.

[11] L. Gatti, and P. Seele, "Evidence for the prevalence of the sustainability concept in European corporate responsibility reporting", Sustainability Science, 9(1), 2014, 89-102.

[12] R. Hahn, and R. Lülfs, "Legitimizing negative aspects in GRI-oriented sustainability reporting: A qualitative analysis of corporate disclosure strategies", Journal of Business Ethics, 123(3), 2014, pp. 401-420.

[13] W.J. Potter, and D. Levine-Donnerstein, "Rethinking validity and reliability in content analysis", Journal of Applied Communication Research, 27(3), 1999, pp. 258-283.

[14] N. Scott, and A.E. Smith, "Use of automated content analysis techniques for event image assessment", Tourism Recreation Research, 30(2), 2005, pp. 87-91.

[15] D.L. Gill, S.J. Dickinson, and A. Scharl, "Communicating sustainability: A web content analysis of North American, Asian and European firms", Journal of Communication Management, 12(3), 2008, pp. 243-262.

[16] R.E. Orwig, H. Chen, and J.F. Nunamaker, "A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output", Journal of the American Society for Information Science, 48(2), 1997, pp. 157-170.

[17] R. Morris, "Computerized content analysis in management research: A demonstration of advantages & limitations", Journal of Management, 20(4), 1994, pp. 903-931.

[18] L. Aaldering, and R. Vliegenthart, "Political leaders and the media. Can we measure political leadership images in newspapers using computer-assisted content analysis?", Quality & Quantity, 2015, pp. 1-35.

[19] N. Péladeau, and C. Stovall, "Application of Provalis Research Corp.'s statistical content analysis text mining to airline safety reports", Retrieved from http://flightsafety.org/files/Provalis text mining report.pdf, 2005.

[20] N. Reuter, S. Vakulenko, J. vom Brocke, S. Debortoli, and O. Müller, "Identifying the role of information systems in achieving energy-related environmental sustainability using text mining", Proceedings of the 22nd European Conference on Information Systems, Tel Aviv, Israel, track 22, 2014, article 12.

[21] J.T. Chang, H. Schütze, and R.B. Altman, "Creating an online dictionary of abbreviations from MEDLINE", Journal of the American Medical Informatics Association, 9(6), 2002, pp. 612-620.

[22] E. Riloff, "Automatically constructing a dictionary for information extraction tasks", Proceedings of the 11th National Conference on Artificial Intelligence, Washington, D.C., 1993 July 11-15, pp. 811-816.

[23] Y. Takayama, Y. Tomiura, K.R. Fleischmann, A.S. Cheng, D.W. Oard, and E. Ishita, "Automatic dictionary extraction and content analysis associated with human values", Information Engineering Express, 1(4), 2015, pp. 107-118.

[24] N. Oostdijk, S. Verberne, and C.H. Koster, "Constructing a broad-coverage lexicon for text mining in the patent domain", Proceedings of the 7th International conference on Language Resources and Evaluation, Valletta, Malta, 2010 May, pp. 17-23.

[25] M.Z. Asghar, A. Khan, S. Ahmad, and B. Ahmad, "Subjectivity lexicon construction for mining drug reviews", Science International, 26(1), 2014, 145.

[26] E. Riloff, and W. Lehnert, "Information extraction as a basis for high-precision text classification", ACM Transactions on Information Systems (TOIS), 12(3), 1994, pp. 296-333.

[27] J.R. Smith, and S.F. Chang, "Searching for images and videos on the world-wide web", IEEE Multimedia Magazine, Retrieved from http://www.ee.columbia.edu/ln/dvmm/publications/96/smith96e.pdf, 1996.

[28] J.R. Park, C. Lu, and L. Marion, "Cataloging professionals in the digital environment: A content analysis of job descriptions", Journal of the American Society for Information Science and Technology, 60(4), 2009, pp. 844-857.

[29] T., Loughran, and B. McDonald, "When is a liability not a liability? Textual analysis, dictionaries, and 10‑Ks", The Journal of Finance, 66(1), 2011, pp. 35-65.

[30] B. de-Miguel-Molina, V. Chirivella-González, and B. García-Ortega, "Corporate philanthropy and community involvement. Analysing companies from France, Germany, the Netherlands and Spain", Quality & Quantity, 2015, pp. 1-26.

[31] J.B. Wade, J.F. Porac, and T.G. Pollock, "Worth, words, and the justification of executive pay", Journal of Organizational Behavior, 18(1), 1997, pp. 641-664.

[32] E. Abrahamson, and M. Eisenman, "Employee-management techniques: Transient fads or trending fashions?", Administrative Science Quarterly, 53(4), 2008, pp. 719-744.

[33] A. Kirilenko, S. Stepchenkova, R. Romsdahl, and K. Mattis, "Computer-assisted analysis of public discourse: A case study of the precautionary principle in the US and UK press", Quality & Quantity, 46(2), 2012, pp. 501-522.

[34] D.N. Bengston, and Z. Xu, "Changing national forest values: A content analysis", Retrieved from http://www.nrs.fs.fed.us/pubs/rp/rp_nc323.pdf. 1995.

[35] C. Lesage, and H. Wechtler, "An inductive typology of auditing research", Contemporary Accounting Research, 29(2), 2012, pp. 487-504.

[36] R. Eriksson, P.B. Jensen, S. Frankild, L.J. Jensen, and S. Brunak, "Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text", Journal of the American Medical Informatics Association, 20(5), 2013, pp. 947-953.

[37] S. Debortoli, O. Müller, and J. vom Brocke, "Comparing business intelligence and big data skills", Business & Information Systems Engineering, 6(5), 2014, pp. 289-300.

[38] A.J. Gill, A. Vasalou, C. Papoutsi, and A.N. Joinson, "Privacy dictionary: A linguistic taxonomy of privacy for content analysis", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2011 May, pp. 3227-3236.

[39] A. Vasalou, A.J. Gill, F. Mazanderani, C. Papoutsi, and A. Joinson, "Privacy dictionary: A new resource for the automated content analysis of privacy", Journal of the American Society for Information Science and Technology, 62(11), 2011, pp. 2095-2105.

[40] Q. Albaugh, J. Sevenans, S. Soroka, and P.J. Loewen, "The automated coding of policy agendas: A dictionary-based approach", *Proceedings of the 6th Annual Comparative Agendas Conference*, Atnwerp, Beligum, 2013 June, pp. 27-29.

[41] R.P. Weber, "Measurement models for content analysis", Quality & Quantity, 17(2), 1983, pp. 127-149.

[42] C.D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, Cambridge University Press, Cambridge, 2008.

[43] J.W. Pennebaker, R.L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015. UT Faculty/Researcher Works", Retrieved from http://hdl.handle.net/2152/31333, 2015.

[44] J. Cohen, "A coefficient of agreement for nominal scales", Educational and Psychosocial Measurement, 20, 1960, pp. 37-46.

[45] J.R. Landis, and G.G. Koch, "The measurement of observer agreement for categorical data", Biometrics, 1977, pp. 159-174.

[46] M. Greenacre, *Correspondence Analysis in Practice* (2nd ed.), Chapman & Hall/CRC, Boca Raton, FL, 2007.

[47] J.J. Inman, V. Shankar, and R. Ferraro, "The roles of channel-category associations and geodemographics in channel patronage", Journal of Marketing, 68, 2004, pp. 51-71.

[48] J. Rojas-Mendez, and M.J. Hine, "South American countries' positioning on personality traits: analysis of 10 national tourism websites", Journal of Vacation Marketing, 2016, Available online first at: http://jvm.sagepub.com/cgi/reprint/13567667166492 27v1.pdf?ijkey=JiURVSvkJcadRB0&keytype=finite

[49] R. Opoku, "Mapping destination personality in cyberspace: An evaluation of country web sites using correspondence analysis", Journal of Internet Commerce, 8, 2009, pp. 70-87.

[50] L.F. Pitt, R. Opoku, M. Hultman, R. Abratt, and S. Spyropoulou, "What I say about myself: Communication of brand personality by African countries", Tourism Management, 28(3), 2007, pp. 835-844.

[51] H. Askell-Williams, and M.J. Lawson, "A correspondence analysis of child-care students' and medical students' knowledge about teaching and learning", International Education Journal, 5(2), 2004, pp. 176-204.