Corporate Social Responsibility Reports: Understanding Topics via Text Mining

Full Paper

Monica Chiarini Tremblay
Florida International University
tremblay@fiu.edu

Carlos M. Parra
Florida International University
cmaparra@fiu.edu

Arturo Castellanos
Florida International University
aancaste@fiu.edu

Abstract
This study utilizes Text Data Mining (TDM) to analyze the contents of Corporate Social Responsibility (CSR) Reports. The goal is to find evidence that environmental sustainability has become embedded in corporate policy and the core business discourse of seven organizations over 2004-2012. Results from supervised modeling techniques suggest embeddedness of environmental qualities in the business discourse. Unsupervised techniques provide additional support for embeddedness—as business topics tend to increasingly group with environmental ones. The process we outline should facilitate pattern discovery in documents, minimizing or eliminating the need for time-consuming content analysis that is frequently used in qualitative research. To our knowledge, this is one of the first attempts to apply TDM processing to analyze unstructured data from CSR reports.

Introduction
Corporate Social Responsibility (CSR) is becoming an increasingly important element of business strategy. Within the overarching CSR integration discussion, there exists a debate around the embeddedness of environmental sustainability in the core business strategy of firms. We investigate whether environmental sustainability remains a concept outside of normal business, or if it has become integrated into business strategy—through firms’ discourse. We utilize Text Data Mining to analyze the content of Corporate Sustainability reports (also referred to as Citizenship Reports, Corporate Social Responsibility Reports, or Corporate Responsibility Reports) produced by seven Dow Jones companies (Citi, Coca-Cola, ExxonMobil, General Motors, Intel, McDonald’s and Microsoft) in three different years: 2004, 2008 and 2012. The 17 reports (seven from 2004, four from 2008, and six from 2012) are downloaded from the corresponding official corporate websites.

TDM adds value to the environmental sustainability embeddedness discussion in at least two ways. First, by demonstrating that environmental sustainability is indeed embedded in corporate policy. Second, unsupervised learning shows clusters that exhibit a tendency for business components to increasingly group with environmental components through time. The process we outline should facilitate pattern discovery in documents, minimizing or eliminating the need for time-consuming content analysis that is frequently used in analyzing CSR reports.

Finding Patterns in CSR Reports
Text categorization can be defined as the assignment of natural language texts to one or more predefined categories based on their content (Dumais et al. 1998). It can be inductive or deductive in purpose and can be used to describe, infer or predict. A commonly used technique to analyze text is content analysis, which requires independent coders to sift through the data and assign parts of the document into predefined codes. These techniques usually involve a manual and time-consuming process (Carley 1993). The evolution of techniques used for CSR Report analysis has advanced from simple word counting to
Analyzing CSR Reports via Text Mining

Text Data Mining (TDM). The first attempts consisted of professionals reading the investment in each CSR area or dimension. Academics have applied content analysis that included manually coding CSR Reports. For example, Moreno and Capriotti (2009) analyzed corporate websites of publicly traded Spanish firms by developing ten content categories and information hierarchies, manually coding, and checking for inter-rater reliability. In their analysis they found that the web has become a prominent medium for communicating CSR issues, however these sites lack external validation for guaranteeing claims. A similar exercise was conducted in the Chinese context by coding newspaper articles on CSR, to characterize the role media should play in facilitating social dialogue (Tang 2012). Another study focused on coding and counting the number of beneficiaries mentioned in CSR Reports, to analyze the impact of CSR initiatives on various quality of life dimensions: income, health, education, markets and democracy (Parra 2008). Finally, and of more interest to the environmental sustainability embeddedness discussion, Ihlen and Roper (2011) adopted an environmental focus to investigate concept use and definitions, as well as the corporate rationales behind sustainability and sustainable development approaches in thirty Fortune 500 non-financial reports in 2006 and 2008. They found that even though environmental issues are increasingly addressed in these reports, the approach is mostly corporate-centric. From the corporate communications perspective, past studies have focused on relating word frequency counts to affective management practices (Saito et al. 2012), brand differentiation (Gill et al. 2008), or have concentrated on a particular industry, such as oil and gas (Dickinson et al. 2008).

Barkemeyer, Figge, Hahn, and Holt (2009) conducted an automated approach on articles published from January 1990 to July 2008. They analyzed 20 million newspaper articles that mentioned to sustainability issues. The goal was to measure the frequency of use of sustainability related terms in media coverage. The researchers utilized stemming algorithms for parsing text in order to identify concepts. They analyzed news articles that referred to Sustainable Development and CSR, rather than Corporate Citizenship or Corporate Sustainability. The next logical extension to this approach is to explore concept associations and knowledge discovery utilizing Text Data Mining (TDM).

Text Mining

Eighty percent of business-relevant information originates in unstructured form, and primarily from text (Grimes 2008). Researchers and practitioners have taken interest in deriving high quality information from text through the use of text analytics and machine learning, or text data mining (TDM). Text mining is a process of knowledge discovery, which allows the organization extract implicit and potentially useful information from textual data using statistical methods (Feldman and Dagan 1995). These algorithms take a statistical approach, calculating word frequencies and term weights to discriminate among a group of documents using similarity detection techniques. High quality in text mining usually refers to some combination of relevance, novelty, and interestingness (Tan 1999) and typical text mining tasks include text categorization, text clustering, concept extraction, sentiment analysis, and document summarization (Dörre et al. 1999).

Text mining algorithms count occurrences of words or phrases in documents. This exercise can quickly grow in complexity as the number of distinct terms across documents increase. As the number of documents increase, the number of columns increase significantly as new terms are added to the document-by-term matrix. A common practice is to use dimensionality reduction techniques to make this matrix compact. For example, algorithms may remove a predefined set of words that have little power in discriminating documents due to its commonality (e.g. the, be, to) (Stop List) or keep only a specific set of context-specific terms (Start List). In addition, to further reduce the sparsity of this matrix we stem words that have a common root (e.g. big: big, bigger, biggest) (Tremblay et al. 2009).

To further improve performance, frequency entries are transformed using weighting schemes (Salton and Buckley 1988). Frequency weights, often called local weights represent the first step in quantifying documents. Unfortunately, absolute counts can be influenced by documents that have high variability with respect to size. For this reason, term weights, often called global weights, modify frequency weights to adjust for document size and term distribution. Research has shown that good results are often obtained using entropy (for short documents) or Inverse Document Frequency (for longer documents, like CSR reports) as the term weight (Woodfield 2011). The log function works best as the frequency weighting with the selection of inverse document frequency. In addition, we leverage on the use of latent Semantic Indexing (LSI) to further reduce the dimensionality using a Singular Value Decomposition.
(SVD). LSI is a technique that transforms the large matrix into a much lower dimensional form to reflect major associative patterns in the data and ignore smaller ones (Deerwester et al. 1990).

In this study we use two approaches for our analysis. First we used a more exploratory (inductive) approach to uncover patterns in the data by analyzing text groupings using unrotated SVDs—similar to the traditional factor analysis with no rotation. The second approach is to use a supervised learning technique using a neural network and memory-based reasoning technique, which is a variation to the nearest neighbor technique that does not require manual topic definitions (Masand et al. 1992).

To our knowledge, this is one of the first attempts that use TDM to analyze the embeddedness of environmental sustainability in unstructured data found in CSR reports. In the subsequent sections of this study we discuss text categorization, our methodology, present our results and analysis, and discuss these results and propose some avenues for future work.

**Methodology**

Using (Gregor and Hevner 2013)’s design science knowledge contribution framework, we position our artifact as an *Exaptation* of existing solutions to a new domain space—that of CSR analysis. IT artifacts can be defined by constructs, models, methods, and instantiations (von Alan et al. 2004). We design and evaluate our instantiation using common performance metrics such as precision, recall, and F-measure. We adopt the framework proposed by Shmueli and Koppius (2011) as an schematic (see Figure 1) of the steps followed in building the explanatory and predictive models in this study.

![Figure 1. Modeling Approach](image)

### Goal Definition

This paper examines the extent to which environmental sustainability has become embedded in corporate policy and core business discourse by analyzing CSR Reports of a sample of large, publicly traded Dow Jones companies. We accomplish so by utilizing TDM to corroborate the work of independent coders and to show how this technique can help streamline content analysis tasks.

### Data Collection and Data Preparation

We obtain text by downloading complete CSR reports in PDF format from the corresponding official corporate websites. In total we download 20 reports from 2004, 2008 and 2012 (Table 1 shows the CSR reports that were downloaded and their size in number of PDF pages). These reports are then manually scrubbed in order to obtain the main text.

<table>
<thead>
<tr>
<th>Company</th>
<th>2004</th>
<th>2008</th>
<th>2012</th>
<th>Average by company</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citi (Citigroup 2006; Citigroup 2009; Citigroup 2013)</td>
<td>56</td>
<td>95**</td>
<td>82</td>
<td><strong>77.7</strong></td>
</tr>
<tr>
<td>Coca-Cola (The Coca Cola Company 2004; The Coca Cola Company 2008; The Coca Cola Company 2012)</td>
<td>44</td>
<td>65</td>
<td>91</td>
<td><strong>66.7</strong></td>
</tr>
<tr>
<td>ExxonMobil (ExxonMobil 2004; ExxonMobil 2008; ExxonMobil 2012)</td>
<td>62</td>
<td>48</td>
<td>67</td>
<td><strong>59</strong></td>
</tr>
<tr>
<td>General Motors (General Motors Corporation 2008; General Motors Corporation 2012)</td>
<td>172</td>
<td>Chapter 11</td>
<td>57</td>
<td><strong>76.3</strong></td>
</tr>
</tbody>
</table>
Table 1. Number of PDF pages in CSR reports for seven U.S. Dow Jones companies in 2004, 2008 and 2012

<table>
<thead>
<tr>
<th>Company</th>
<th>2004</th>
<th>2008</th>
<th>2012</th>
<th>Average by year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel (Intel 2004; Intel 2008; Intel 2012)</td>
<td>40</td>
<td>108</td>
<td>126</td>
<td>91.3</td>
</tr>
<tr>
<td>McDonald's (Corporation 2008; Corporation 2012)</td>
<td>88</td>
<td>70</td>
<td>8*</td>
<td>55.3</td>
</tr>
<tr>
<td>Microsoft (Microsoft Corporation 2004; Microsoft Corporation 2008; Microsoft Corporation 2012)</td>
<td>80</td>
<td>5*</td>
<td>89</td>
<td>85</td>
</tr>
</tbody>
</table>

**Table 1. Number of PDF pages in CSR reports for seven U.S. Dow Jones companies in 2004, 2008 and 2012**

* Document too small to be partitioned and analyzed
** Un-editable file could not be scrubbed or partitioned

We continue by partitioning each report into 5 CSR dimensions that are based on the Sustainability Accounting Standards Board (SASB) (2014). This guideline contains 5 dimensions: a business partition, a governance partition, an environmental partition, a human capital partition, and a social capital partition. Each file is partitioned and classified by a subject matter expert. Not all of the 20 reports downloaded were used in the analysis. Citi’s 2008 CSR report had security settings that did not allow us to obtain main text, while Microsoft’s 2008 and McDonalds’ 2012 CSR reports were too short to obtain meaningful partitions. Thus, out of the 17 reports used in the analysis a total of 85 partitions were obtained - 35 partitions from 7 reports for 2004; 20 partitions from 4 reports for 2008; and 30 partitions from 6 reports for 2012.

**Exploratory Data Analysis and Choice of Variables**

To reiterate, text mining takes all the words found in the input documents, indexes, and counts them in order to compute the term-document matrix. This process can be refined to exclude certain common words such as "the" and "a" (stop word lists). We run the unsupervised TDM on 85 partitions to identify words that do not add value to term groupings and that should be included in the stop list (see Table 2). For example, consider the words “Intel” or “Microprocessor” as terms with high weights associated to Intel’s 2004 social capital partition. These words may be considered of a higher weight because they appear frequently. However, they do not add value and could overshadow less frequent yet important words, such as “teacher”. Thus, we included all these non-value-added words in the following list:
Table 2. List of words included in Stop List

As part of our exploratory analysis we created a profile of the most descriptive terms—terms that had a higher weight—for each of the target categories (See Table 3).

Table 3. Descriptive terms for each of the target categories

<table>
<thead>
<tr>
<th>Target</th>
<th>Descriptive Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business</td>
<td>product, customer, choice, financial, child, food, demand, technology</td>
</tr>
<tr>
<td>Environment</td>
<td>environmental, water, water, package, facility, reduce, energy, emission</td>
</tr>
<tr>
<td>Governance</td>
<td>code, policy, board, compliance, director, corporate committee, company</td>
</tr>
<tr>
<td>Human Capital</td>
<td>employees, right, diversity, workplace, human, human right, workforce, aids</td>
</tr>
<tr>
<td>Social Capital</td>
<td>supplier, local, education, woman, community, foundation, program, support</td>
</tr>
</tbody>
</table>

Choice of Methods

Data mining tasks can be broadly categorized into either supervised or unsupervised learning tasks. In supervised learning, the classifications are assigned and are known before computation based on a “gold standard”. In unsupervised learning, datasets are assigned to segments without knowing—a priori—the classification. In this study we utilize both techniques for two different purposes. We utilize supervised TDM on the whole corpus to verify whether deductive machine learning would perform in a similar way as humans at the task of classifying contents from CSR reports into one of the five Sustainability dimensions. If this was the case, we could then use unsupervised TDM to look at how these partitions group and explore whether environmental sustainability considerations are part of the core business discourse for the firms analyzed.

Supervised TDM

For the supervised component of this study we use the labels assigned by the subject matter expert, such that each text file in the input data is associated to a partition type. Input data are divided into a training set and a validation set. In essence, 60% of the documents in the input data are used to train several machine learning algorithms. The remaining 40% are used to validate and test the performance of the text
classification models by having it assign labels to the remaining documents—the set of files not used during training.

We train several models to compare the classification accuracy of several algorithms; this allows us to select the best model for the task. We describe the results of our three best models: a decision tree model, a neural network model, and a memory-based reasoning model. We compare this models using common standard metrics such as accuracy, precision, recall, F-measure, sensitivity, and specificity (Jurafsky and Martin 2009) (see Tables 4 and 5). The accuracy of a classifier is calculated by dividing the correctly classified instances by the total number of instances. Recall (R) reflects the percentage of correct positive predictions out of all the possible positives in the evaluation set, while precision (P) reflects the percentage of correct positive predictions out of the predicted positives. The F-measure is simply a ratio of overall goodness of fit for precision and recall. Sensitivity (TPR) measures the proportion of actual positives that are correctly classified as positive, whereas specificity (SPC) measures the proportion of actual negatives that are correctly identified as negatives. In addition, we refer to misclassification rate as the number of incorrect classification over the total number of classifications.

\[
\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F = \frac{2(P \cdot R)}{P + R}, \quad \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
\]

Table 4. Formulas for Evaluation

Since we are dealing with a multi-class problem in which the target variable can have five different levels—business, environment, governance, human capital, or social capital—we use a weighted average of these individual measures. We did this by calculating the macro-average and the micro-average (Yang 1999). Macroaveraging gives equal weight to each class and gives a sense of effectiveness on small classes, whereas microaveraging gives equal weight to each per-document classification decision (Manning et al. 2008). We calculate the macro-average and micro-average precision, recall, and F-measure and we report the macro-average measure. On average, the absolute difference between both calculations did not exceed two percent.

Table 5. Evaluation metrics for the three supervised models

Our results indicate that the supervised learning classifications are similar to that of the subject matter expert. The Neural Network model had the lowest misclassification rate—and of the 40 sustainability partitions in the validation set, only four were classified differently (“misclassified”). These misclassified partitions were initially labeled as business partitions (ExxonMobil 2008; ExxonMobil 2012) and as social capital partitions (General Motors Corporation 2012; Intel 2012) and were classified by the algorithm as environmental partitions. All other partitions were given identical labels, agreeing with the work done by the subject matter expert.

The Memory-based Reasoning algorithm had five misclassifications—a misclassification rate of 12.5%. Three of these were business partitions (ExxonMobil 2008 and 2012, as well as Coca-Cola 2012 business partitions (ExxonMobil 2008; ExxonMobil 2012; The Coca Cola Company 2012) that were classified as environmental ones. The remaining two were social capital partitions (ExxonMobil 2012; General Motors Corporation 2012). General Motor’s 2012 social capital partition was reclassified as an environmental partition, while ExxonMobil’s 2012 social capital partition was reclassified as a human capital partition. Finally, the Decision Tree algorithm misclassified seven documents. We do not explain these in detail due to its high misclassification rate—almost 18%.
Our results show that supervised learning techniques can accurately classify documents based on the Sustainability Accounting Standards Board (SASB) framework that was trained on. We now investigate the use of unsupervised learning techniques to explore alternative document classifications.

Unsupervised Learning Strategy

Text cluster analysis encompasses a number of different classification algorithms that can be used to develop taxonomies (typically as part of an exploratory data analysis). The expectation-maximization algorithm used relies on the singular-value decomposition (SVD), which transforms the original term-document matrix into a dense, low-dimensional representation in order to identify clusters in the data that are not pre-determined by a label (Dempster et al. 1977; Do and Batzoglou 2008; Zhang et al. 2005).

The unsupervised learning strategy was applied to each year separately in order to perform a longitudinal analysis of the discourse evolution over the three time periods. For each of the time periods (2004, 2008, and 2012) we show the document groupings and a table that includes the cluster name, the descriptive terms of the cluster, and which documents fall under each of the clusters (Parra and Tremblay 2014).

In 2004, the text cluster analysis resulted in six groupings (see Figure 2): human and social capital; environment; business and social capital for Tech firms; business and governance for Hydrocarbon firms (oil extraction and automotive); financial firm (Citi); and governance grouping (see Table 7).

2004 Document Groupings

![Figure No. 2, Text Cluster analysis on 2004 partitions](chart)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster Name</th>
<th>Descriptive Terms</th>
<th>Files Grouped</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Human and Social Capital</td>
<td>health, diversity, skills, education, safety, training, related, annual, +focus, national, +review, benefits, +place, best, +service</td>
<td>Citi2004: governance, human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coke2004: human capital, social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExxonMobil2004: human capital, social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>General Motors2004: human capital, social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intel2004: human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>McDonalds2004: human capital</td>
</tr>
</tbody>
</table>
Table 7. 2004 Document Groupings Description

In 2008, the cluster analysis suggested three different groupings: business for food and beverage firms; human capital, social capital, and governance; and the environment grouping (see Figure 3). One of the business documents—Exxon-Mobil—was grouped into the Environment cluster. Upon further analysis of this particular document, we noticed it referred to Exxon-Mobil’s challenge of keeping up with the increasing energy demand while reducing the environmental impact. This particular report focused in improving the efficiency of Exxon-Mobil’s operations and products, “while developing technologies that can substantially reduce the environmental footprint of energy use for the long term”, implicitly referring to terms that were part of the environment grouping (ExxonMobil 2009). The SASB categories were not explicitly addressed in these documents—topics were intertwined with one another with no clear distinction across the different categories (see Table 8).
### 2008 Document Groupings

![Figure 3. Text Cluster analysis on 2008 partitions](image)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster Name</th>
<th>Descriptive Terms</th>
<th>Files Grouped</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Business for Food and Beverage Firms</td>
<td>+brand, children, +act, +good, conduct, values, independent, markets, quality, +set, guidelines, corporate, specific, local, +control</td>
<td>Coke2008: business, governance, social capital</td>
</tr>
</tbody>
</table>

Table 8. 2008 Document Groupings Description
In 2012, the analysis revealed six different groupings: oil extraction firm, environmental, governance, social capital, beverage firm, and human capital grouping (see Figure 4). Interestingly, General Motors’ 2012 and Intel’s 2012 business partitions were grouped together in the environmental cluster (No. 2) when using unsupervised learning techniques. This indicates that in these business partitions environmental sustainability considerations were heavily weighted. General Motors produces cars, which consume fuel and contribute to greenhouse gas emissions thus the importance of reducing environmental footprint in their core business discourse; and Intel’s goal of “designing products with improved energy-efficient performance” helps reduce emissions and energy costs. Thus, environmental sustainability considerations are likely included in their core business discussion (see Table 9).

2012 Document Groupings

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Cluster Name</th>
<th>Description</th>
<th>Files Grouped</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ExxonMobil</td>
<td>Grouping points to <strong>ExxonMobil considerations</strong>, this cluster groups all ExxonMobil partitions except governance and points to firm specific issues such as: upstream lines, gas, plans, projects, operations, conditions, international, design.</td>
<td>ExxonMobil2012: business, environment, human capital, social capital</td>
</tr>
<tr>
<td>2</td>
<td>Environment</td>
<td>Grouping refers to <strong>environmental</strong> issues (i.e. air, carbon, power, waste, drive, water, cost, energy, etc.). It is important to note that General Motors and Intel business partitions also grouped here, indicating that the descriptive terms are increasingly relevant for these firms’ business activities.</td>
<td>Citi2012: environment, GeneralMotors2012: business, environment, Intel2012: business, environment, Microsoft2012:</td>
</tr>
<tr>
<td>#</td>
<td>Grouping</td>
<td>Description</td>
<td>Example Firms</td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>3</td>
<td>Governance</td>
<td>Grouping includes most governance partitions and mentions: conduct, directors, compensation, independent, legal, principles, review, compliance, risk, etc. Citi’s business and social capital partitions also grouped here, evidencing that governance topics were a priority for Citi in 2012.</td>
<td>Citi2012: business, governance, social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coke2012: governance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ExxonMobil2012: governance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GeneralMotors2012: governance</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Microsoft2012: governance</td>
</tr>
<tr>
<td>4</td>
<td>Social Capital for Tech Firms</td>
<td>Grouping brings up social capital considerations for Tech firms insofar as it groups Microsoft and Intel partitions alluding supply chain issues (i.e. cash, central, campaign, audits, future, serve, conditions, china, centers, etc.).</td>
<td>Intel2012: social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Microsoft2012: social capital</td>
</tr>
<tr>
<td>5</td>
<td>Coca-Cola</td>
<td>Grouping includes most of Coca-Cola’s partitions (business, environment and social capital) – as well as General Motors social capital partition - and includes the following terms: “grant, Brazil, water, fund, partner, china, campaign, waste, partners, etc.” Thus, this Text Cluster raises Coca-Cola considerations.</td>
<td>Coke2012: business, environment, social capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GeneralMotors2012: social capital</td>
</tr>
<tr>
<td>6</td>
<td>Human Capital</td>
<td>Finally, this grouping points to human capital considerations across industries through: “top, culture, women, workplace, safety, directors, people, human rights, etc.”</td>
<td>Citi2012: human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Coke2012: human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>GeneralMotors2012: human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Intel2012: governance, human capital</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Microsoft2012: business, human capital</td>
</tr>
</tbody>
</table>

Table 9. 2012 Document Groupings Description

In summary, 2004 has more industry specific groupings (for Tech firms and for Hydrocarbon firms), in 2008 there was only one industry specific grouping (for food and beverage firms) and in 2012 individual firm groupings gained more relevance (ExxonMobil grouping and a Coca-Cola grouping). This indicates that groupings have become less sector specific but more related to individual firms. Even though there are more industry groups guiding firm behavior, even in CSR terms, firms want to differentiate themselves from their competitors. In 2004, the environmental grouping did not include any business partitions; in 2008, it included one business partition (from ExxonMobil); and in 2012, the environmental partition also grouped two business partitions (from General Motors and Intel). Thus, business partitions tend to increasingly group with environmental ones indicating higher integration of environmental considerations in business activities. Finally, in 2004 and 2012, the cluster analysis...
produced a standalone governance grouping, which was not apparent in 2008—except for its implicit inclusion in the food and beverage grouping.

**Conclusion and Future Work**

In this study we illustrate how researchers can leverage the use of supervised TDM techniques to automate content analysis in large documents with unstructured data. We found that TDM adds value to the environmental sustainability embeddedness discussion. Using supervised learning techniques—we are able to show how the core business discourse incorporates environmental sustainability considerations. We ran several algorithms, and since we are dealing with a multi-class problem in which the target variable can have five different possibilities, we used a weighted average of individual accuracy measures. We did this by calculating macro-average and micro-average precision, recall, and F-measure. The best performing algorithm was Neural Network, assessed by the misclassification rate, which showed that CSR report partitions focusing on core business discussion (in the case of ExxonMobil), and social capital topics (in the case of General Motors), contained enough environmental sustainability considerations to be classified as environmental partitions as opposed to the original classification, which included them in the business category. Our results show that supervised learning techniques can accurately classify documents based on the Sustainability Accounting Standards Board (SASB) framework that it was trained on.

Second, we used unsupervised learning techniques (cluster analysis) results in clusters that exhibit a tendency for business components to increasingly group with environmental components through time. In 2004, the environmental grouping did not include any business partitions, but in 2008 it included one business partition (from ExxonMobil), and in 2012 the environmental cluster grouped two business partitions (from General Motors and Intel). Thus, business partitions tend to increasingly group with environmental ones indicating higher integration of environmental considerations in business activities.

Finally, it makes sense for companies like ExxonMobil and General Motors to have their business and social capital partitions be reclassified by supervised TDM as environmental ones. And also, unsupervised learning shows their business partitions grouping in the environmental cluster. For us this portrays a logical evolution for firms whose products both consume natural resources and contribute to global warming.

This study is not without limitations; in our analysis we only consider a small sample of large, publicly traded Dow Jones companies across three different time periods. As a result, future work should focus in adding more organizations and perhaps extend the analysis by segmenting the organizations across different sectors (e.g., healthcare delivery, retail, apparel). The results of this analysis should help the organization assess the impact of their embeddedness efforts and strategize accordingly. Moreover, future work should study how embeddedness relates to an organization’s financial performance. Finally, we need to investigate the scalability of text data mining to a larger or different context and a larger amount of documents with less human labor.

**REFERENCES**


Analyzing CSR Reports via Text Mining


Ihlen, Ø., and Roper, J. 2011. "Corporate Reports on Sustainability and Sustainable Development:‘We Have Arrived’: Sustainable development)."


Woodfield, T. 2011. Text Analytics Using Sas Text Miner Course Notes. Cary, North Carolina: