AUTOMATIC ANALYSIS OF CORPORATE SUSTAINABILITY REPORTS AND INTELLIGENT SCORING

AMIR MOHAMMAD SHAHI

Faculty of Engineering, Computing and Science, Swinburne University of Technology (Sarawak Campus)
Kuching, Sarawak, Malaysia
amir@amirms.com

BIJU ISSAC

School of Computing, Teesside University
Middlesbrough, UK
b.issac@tees.ac.uk

JASHUA RAJESH MODAPOTHALA

School of Business, Monash University
Bandar Sunway, Malaysia
jashua.rajesh@monash.edu

As more and more corporations and business entities have been publishing corporate sustainability reports, the current manual process of analyzing the reports is becoming obsolete and tedious. Development of an intelligent software tool to perform the report analysis task would be an ideal solution to this long standing problem. In this paper we argue that, given sufficient quality training using a custom corpus, corporate sustainability reports can be analysed in mass numbers using a supervised learning based text mining software. We also discuss our methodologies of improving the accuracy of our classifier as well as the feature selector in order to gain better performance and more stability. Additionally, the achieved results of executing the developed software on one hundred reports are discussed in order to prove our claims.

Keywords : machine learning; text analysis; text mining; clustering; classification; association rules

1. Introduction

Companies, corporations and businesses produce and publish various types of qualitative documents periodically. This is provided to demonstrate their different aspects of business related to their current and future stakeholders and to improve their overall management and leadership image. The number of these reports sums up to hundreds of thousands around the world every year; this has motivated many international organizations such as Global Reporting Initiatives (GRI) to develop their publicly available guidelines on how to collect, create and rate such reports. Complying with the well-known standards brings additional credit to the reporting organization and is the main reason behind the growing trend in number of consultancy firms, which help businesses to get their reports analysed, fine-tuned and certified.

Currently, the process of checking reports; compliance with standard guidelines is done manually by teams of domain experts and certified consultants. This is considered to be time and resource intensive, daunting and human error prone.

The growing number of corporate sustainability reports (CSR) publications has created an overwhelming demand for analysis and pre-scoring of such reports before being submitted to certification authorities. The cost of such analysis is however too high for many organizations to afford as it is currently conducted manually by a team of experts. Availability of high performance computer hardware along with the current

NOTE: The preprint of an article published in [International Journal of Computational Intelligence and Applications (IJCIA), Vol 13, Issue 1, 2014, 27 pages]; [Article DOI: 10.1142/S1469026814500060]

sophisticated software technologies enables building an automatic report analysis software package within reach and highly desirable.

Due to the large amount of data and information involved in Corporate Sustainability Reporting, there has been always a considerable demand for implementing automated solutions for both generating and analysis of such reports [1, 2].

Extensive attention has been paid to development of technological infrastructure and tools for automation of generating CSR reports [3, 4], but developing an automated CSR report analysis system has been widely overlooked by the research community. Development of such system had been anticipated since the early days of GRI 3.0 [5] and it has been called to be highly desirable [6]. In this research, we have carried out an investigation to find out the possibility of measuring the completeness of GRI Corporate Sustainability Reports i.e. assessing them based on GRI 3.0 Content Index to find out the sections, which fulfil particular Performance Indicators. This will help the report analysis system figure out whether or not a CSR report complies with the official guidelines of report completeness. In this research, we treated each section of the CSR reports as a document, which was expected to fall under one of the Performance Indicators, considered as a category.

As the software shall be able to perform such document categorization automatically, we propose using the Machine Learning approach to Text Categorization (TC) in a supervised-learning environment. This is due to the proven high effectiveness and relatively low costs of such approach [7]. This paper shows the results of our research to illustrate the suitability of Supervised Machine Learning in development of such tool. Our evaluations showed that Naïve Bayes and Decision Tree algorithms produce the best results among other learning methods. Furthermore, we investigated various methods of classification optimization among which we chose to combine a correlation-based feature selection algorithm with some of our classifiers; the combination made the learning process gentler and yielded more accurate results. Lastly, we scored 100 reports against the G3 framework using our developed software and compared the results with actual report scores. The results of the comparison are also reported in this paper.

The structure of this paper is as follows. Following this introduction section the paper reports on the related works and technologies in the field in Section 2. The third section discusses the details of our conducted experiment such as the applied methodology and research parameters as well as brief presentation of results achieved in each development phase. The overall results of execution achieved after development conclusion are thoroughly discussed in Section 4. In order to provide a sense of applicability of the developed system in business context, we executed our developed system on real-world CSR reports; the results of this experiment along with some suggestion on other possible experiments are discussed in Section 5 followed by the conclusion section in which the findings of the research are briefly discussed and certain improvement suggestions are made.

2. Related Works and Technologies

2.1. *Machine Learning in Text Mining*

Text mining, also referred to as Text Analytics, is the process of extracting useful information from textual data through analytical methods of data mining such as statistical pattern learning. It can be seen as a data mining technique the input of which is of natural language text; By making this assumption, one could easily figure out that text mining systems shall conduct their task by using text-specific techniques such as lexical analysis, word frequency statistical analysis, Natural Language Processing (NLP) as well as the generic data mining techniques.

Although commercial-grade text mining systems have been emerging only in the past 10-15 years, computational linguistics has been a topic of research for decades perhaps initiated by Weizenbaum's Eliza system in the early days of artificial intelligence [8], which applied basic pattern matching and linguistic rules to mimic a psychotherapist in a dialogue with a patient.

It is argued that over 80% of business-worth information is locked in unstructured textual form [9]; this makes text mining have extreme commercial value, which can be used in solving various text analysis problems such as text clustering, taxonomy extraction, sentiment analysis, text summarization as well as text categorization.

2.1.1. *Text Mining in Text Categorization*

Text Categorization generally and the Machine Learning approach to it specifically have been subject to research aiming at solving various document analysis problems since their early introduction. Textual documents are generally categorized by their attributes such as author, publication year, publisher and subject. Categorizing textual documents based on their subject is a prominent problem as the volume of text to be analysed for subject selection is enormous compared to amount of text of other attributes. The high volume of text besides our limited capability of performing repetitive tasks makes textual document categorization by subject an extremely difficult task to be done manually and therefore an interesting problem to be solved by computers [7]. In fact, Most of the research has been conducted on those problems, which being otherwise solved by manual means would be either too difficult, expensive or even infeasible. These problems include (but not limited to):

(1) Document Organization: This would involve analysing the contents of documents such as news articles to be classified under a predefined set of categories such as sports, politics, society, etc. [10]

(2) Text Filtering: Implemented on either sender or receiver side involves determining whether a given document is suitable to be sent/received [11, 12]. A good example is the news corporations which produce news for their broadcasting clients. A client who publishes news regarding sports would not be interested in receiving news of any other topics, therefore the streamed news articles to this client shall be filtered at either the sender side or the client side.

(3) Patent application categorization: Involves filing submitted patent applications under their respective categories, by studying the most important parts of the application such as the title and the first and last clauses, date of submission and applicant's name [13, 14].
(4) Spam Detection: Perform Boolean classification of incoming messages by investigation their different components such as header, body, meta-data, etc. to categorize them as either Spam or Not Spam [15-17].

As many as 50 different document classification algorithms have been implemented in WEKA data mining and machine learning library [18]. These classifiers are of different types such as Probabilistic, Decision Tree, Decision Rule, Regression Method, Online Method, Batch Linear Methods, Neural Networks, Example Based and Support Vector Machines. Since little research has been done on conducting an effectiveness comparison of all of them in a controlled environment, it is very difficult to choose the most appropriate one for a given problem. An environment is considered being "Controlled" if and only if the tests conducted in it have been be done by a single author under similar conditions [7]. Sebastiani has ranked the classifiers based on their reported relative performance on similar datasets. They found the Support Vector Machines, Example Based Methods and Regression Methods to provide the best performance followed by Probabilistic and Batch Linear Methods [7]. However, they were unable to measure the performance of Decision Table and Rule Induction algorithms due to lack of sufficient literature and results at time of their research.

A more recent study by Shen, et al. [19] which compares the classification accuracy of nine classifier algorithms in relation to prediction accuracy of liver cancer of 88 test cases (59 with liver cancer and 29 without cancer) concludes that Support Vector Machines with radial kernel features the most accurate classification model at approximately 67% true positive detection rate. However, we must bear in mind that the accuracy of classifier algorithms can differ from a problem case to the next and we should not rely solely on findings of other researchers when choosing the most appropriate sort of algorithm(s) to solve a new problem [20].

2.1.2. *Text Mining in Business Content Analysis*

Wilson and Rayson [21] believe that content analysis is a form of quantitative research, but it is different from traditional quantitative research because it deals with free text which has not been collected using a pre-coded questionnaire. They argue that the main concern of content analysis is 'statistical analysis of primarily the semantic features of texts' i.e. categorizing sections of texts under a given set of categories.

Computer aided content analysis dates back to 1960's which was proven feasible by the sophisticated General Inquirer software of Harvard University [22]. Ever since General Inquirer was developed, there have been many research efforts into applying machine intelligence in textual content analysis and many software tools have been developed for this purpose [see 23], many of which have produced fascinating results. The impressive work by Crossley and McNamara [24] who successfully applied supervised learning in discerning patterns related with text patterns of native and non-

native English writers is a good example worth looking into. They managed to develop a learner which distinguishes the level of English skills between the English essays written by English-speaking students (L1) and Spanish-speaking students (L2) based on their belief that 'L2 writers of English differ from L1 writers in their use of lexical cohesive devices and other lexical features'. I am also impressed by the work of Wilson and Rayson [21] who reported on their attempt in developing the Lancaster Content Analyser which uses Natural Language Processing techniques to extract rule sets from a preliminary corpus in order to assign semantic tags to large bodies of transcribed spoken interviews between members of the public and market researchers. Their developed prototype system was reported to produce a success rate of over 90%, which is very striking although no further report has been published on its further developments.

Machine intelligence has also been subject to research into similar domains such as metaphor analysis [25], language translation studies [26],cross-lingual semantic tagging [27], keyword extraction from full text [28]. McDonald, et al. [29] also found text mining research projects in biomedical sciences, chemistry and some early adoptions in social sciences and humanities.

Text mining techniques have been contributing a positive impact on business by discovering hidden, undiscovered and overlooked data patterns in various business data resources such as blogs, websites, social media contents, etc. for business intelligence purposes –such as discovering business trends and customer preferences. The extracted information is often used for gaining competitive edge by providing newer and better services and products or for research and development reasons. In his Master's thesis, Herron [30] pointed out that scholarly articles and patents are currently mined by the pharmaceutical industry in order to discover drug usage trends and possible drug alternatives.

While text mining is rapidly becoming a major revenue stream for many companies, from the well-known giants such as IBM and Oracle to smaller companies such as ScrapperWiki and SAS, governments and security agencies are also making considerable investments in the field for various purposes such as legal case analysis [31] and counter terrorism [32]. After all, with the exponential growth of data production, predicted to be at a 40% p.a. rate, artificial intelligence based data exploration solutions such as text mining have significant potential societal and economic value [29].

2.2. *Automatic Analysis of Corporate Sustainability Reports*

2.2.1. *Background*

Reporting on corporate sustainability performance has been gaining popularity as businesses have been showing increasing interest in reporting on not only their environmental performance, but also their economic and social performance in an integrated report [33]. According to GRI reporting statistics, more than 1800 businesses have produced and published their CSR reports in 2010 from which 125 are among the European Union's Global 500 companies [34]. Furthermore, a 2011 survey by KPMG found that nearly 95% of the largest 250 companies in the world publish CSR reports

[35].This wide and popular interest in reporting on sustainability is witnessed despite the fact that doing so is completely voluntary in most countries and very costly [36-38]

Systematic research has been conducted into the reason behind such increasing interest by many scholars. Various internal and external forces and motivations are believed to drive the exponential growth. Azzone, et al. [39] believe that the main objective of environmental reporting is to communicate environmental performance, acknowledgement of environmental responsibility, gaining competitive edge, obtaining social approval and showing regulatory compliance. Pressure from local governments and legislations are also believed to be a major influencing factor [33]. Sumiani, et al. [36] have gone further by splitting such intentions into 1) Motivations behind reporting on social performance and 2) Pressing forces behind reporting on environmental performance. They argue that factors such as economic and market pressure, environmental crises and high population growth rate motivate managers to report on their business's social performance while informing and benefiting stakeholders, pressure from various interest groups and political and cultural conditions of the host country force them to consider reporting on their environmental impacts.

The explosion in number of published reports as well as the number of pages per report [40] signifies the need for report quality metrics i.e. mechanisms to measure the quality as well as the completeness of CSR reports [41]. We need such metrics to be able to monitor corporations' advancement toward sustainable development as argued by Hussey, et al. [42] who studied CSR reports published between 1995 to 2000 to conclude that a commonly accepted metric helps corporations to gauge improvement, impact consumer vote and influence regulatory action.

While the validity of claims made in CSR reports –or 'quality' of reports, cannot be assured by reading the reports alone [43], it would be possible to score the reports based on their 'completeness' if measured against an indicator-based reporting framework. An indicator-based framework is the ideal solution for CSR reporting as it simplifies the scoring process and provides a common language for complex issues [44]. Need for an indicator based reporting framework was sensed in the 1990's due to the information explosion phenomena and at the same time, firms are showing increasing interest in complying with international reporting framework to ease external validation processes [33]. These factors have made many international organizations develop such frameworks.

Since the release of the third version of GRI's guideline, also known as GRI 3, it has become the de-facto standard framework for corporate sustainability reporting [2, 45]. One of the most important reasons behind such warm acceptance of the framework is its comprehensiveness in covering almost all of generic social, environmental and economic aspects of sustainable development [42]. To address the specific needs of certain business domains and industries also highlighted by Scott Marshall and Brown [38], GRI has recently developed supplementary kits for some domains to address those specific reporting needs. It is also widely believed that GRI accredited firms have higher sustainability performance than those who use other reporting frameworks [35]. Published reports are later submitted to either a self-hired consultancy firm or GRI

organization to be given a score, which exposes to public the company's environmental, social and governance performance [46].

2.2.2. *Intelligent Approach to CSR report analysis*

As the number of published reports is increasing exponentially, one could easily see the need for an intelligent software system to help reporting entities and report assurance firms with scoring corporate sustainability reports. Due to the large amount of data and information involved in Corporate Sustainability Reporting, there has been always a considerable demand for implementing automated solutions for generating and analysing such reports [1, 2].

Extensive attention has been paid to development of technological infrastructure and tools for automation of generating CSR reports [3, 4] and to development of newer CSR publishing methods [37, 45, 47], but developing an automated CSR report analysis system has been widely overlooked by the research community even though development of such system had been anticipated since the early days of GRI 3.0 [5] and it is called to be highly desirable [6, 48].

Measuring the 'completeness' of CSR reports, taking into account their qualitative, general and highly descriptive nature [49], is a daunting, resource intensive task specially when report comparison is to be undertaken [50].

Our search for prior research efforts in applying machine learning and data mining for CSR report analysis and scoring yielded no significant results except for the works by Modapothala and Issac who had successfully taken this approach to discover the reporting patterns across various industries and business domains [See 48, 50, 51, 52]; Although they made use of very few variables in their analysis when compared to a sophisticated machine learning technique and their works were not aimed at scoring the reports as per GRI application level guidelines, they shed a good light on suitability of applying data mining techniques for other CSR report analysis goals e.g. report scoring.

Nevertheless, some other prior efforts on analysis of reports, other than CSR reports, using text mining have been reported on. Among the most recent is the interesting research by Botsis, et al. [53] who developed a text mining system, called VaeTM, using which they managed to extract primary (diagnosis and cause of death) and secondary features (e.g., symptoms) from hundreds of vaccine adverse event reporting system (VAERS) reports. Their text miner yielded an encouraging 83.1% effectiveness which is two times more effective than other comparable tools available online. Eckstein [54] has also reported on developing a machine learning based system which uses Naïve Bayes and Support Vector Machines to analyse thousands of outbreak reports aiming at identifying the nosocomial outbreaks (i.e., outbreaks in hospitals and other health care facilities). To name an even more interesting project one may want to point at making company bankruptcy predictions by analysing the qualitative sections of corporate financial reports. In their project, Shirata, et al. [55] developed a text mining system using which they analysed hundreds of corporate financial reports believing that it would be easier to notice signs of financial positions in nonfinancial information than in financial figures. The text mining tool successfully identified certain nonfinancial key phrases

appearing which in financial reports indicate predictable bankruptcy. Prasad, et al. [56] developed a preliminary text mining system which examines free text radiology reports in order to convert them to structured XML reports. A similar effort by Friedlin, et al. [57] introduced a medical report analysis framework using an annotated semantic index. Both teams report promising results and agree that using machine learning based approach to text processing is the select approach to free text report analysis.

3. Details of Experminent

This research was conducted in three (3) stages; initially, the effectiveness of various machine learning algorithms were tested in order to select the best performing ones. This was done by conducting a train-and-test effectiveness testing on a training corpus we had created earlier. The corpus contains thousands of training samples we had extracted from actual CSR reports. Later after choosing the top performing classifier algorithms, we attempted to boost their performance, in both classification speed and effectiveness. Lastly, the software was implemented using the optimized algorithms to score 100 CSR reports the actual scores of which were compared with software determined scores to measure the software's accuracy. Details of our three-stage experiment follow.

3.1. *Stage 1: Classifier Algorithm Benchmark*

3.1.1. *Methodology*

We carried out our research by creating a training corpus for our machine learning algorithms to initiate the machine training on. At the final stage, we tested different document classification algorithms on the testing set to identify the methods with the highest accuracy. Our approach to confirming reliable results is therefore the train-and-test approach as mentioned in [7].

Preprocessing:

In order to avoid the common problem of Curse of Dimensionality [7, 58-61], we reduced the dimensionality of our corpus by filtering out the usual English stop words. Numbers, qualifiers, pronouns, prepositions, adjectives and adverbs were also filtered out as suggested in [60]. The same process was iteratively applied to any future document before being classified, thus converted the document to a vector of terms $T = \{T1,\ldots,T|t|\}$. This implies the feature extraction approach to dimensionality reduction [58].

Inductive Training and Testing of Classifiers:

We selected 4 document classification algorithms to be trained on our corpus. Some of them produce the most top-notch results in equal environments [7]. We selected the following classifiers to be studied upon:

- Naïve Bayes Classifier: Naïve Bayes is a probabilistic classifier combining the Bayes' Theorem with some basic (naïve) independence assumptions such as total independence of document features.

- Decision Table: Decision Table has two main components: 1) a set of decision features, called Schema and 2) a document space called Body, consisting of labelled documents from the document space defined by the Schema. Classification of an unclassified document is performed by attempting to find an exact match of its features in the Schema. If none is found, the majority class is returned [62].
- Random Sub Space: Random Subspace is a Decision Tree Based classifier, which attempts to improve the overall generalization accuracy while maintaining the highest accuracy on training data. Combining multiple randomly created trees is the main characteristic of this classifier. The trees are constructed systematically by pseudo randomly selecting subsets of feature space [63].
- Neural Networks: An artificial intelligence method based on interconnection of artificial Neurons. It keeps searching for the optimal solution while it can improve the quality of the current network. Eventually, It returns the most suitable network as the result [64].

In order to test the effectiveness of each classifier, we split the training corpus in two sets:

A training set TrS = {d1,…,dTrS} on which each classifier Q was built through receiving training i.e. by observing the characteristics of each document classified under a category C = {c1,…,c|c|}.

A testing set TtS = {dTrS+1,…,d|d|} which was used to test the effectiveness of classifiers. Each document dj would be fed to the classifier Q for its decision on Q (di,cj) to be testified against that of a domain expert i.e. $\bar{Q}$ (dj,ci). The effectiveness of the classifier is based on how often the classifier decisions and the expert decisions match.

The training and testing set were randomly populated with sample documents on each execution.

3.1.2. *Research Parameters*

In order to benchmark the performances, we conducted our experiments by setting up the following experimental environment.

Training Corpus:

As an integral part of the Supervised Learning approach [7, 60], we trained our system inductively on how to classify textual documents based on the characteristics observed in sample training documents also known as training corpus. The corpus contains actual text from real world CSR reports categorized under appropriate categories (Performance Indicators) by either GRI Organization or third party firms. We selected the Environmental subclass of GRI 3.0 Content Index and each of its Performance Indicators was treated as an individual, independent, and mutually exclusive category. Selected parts of the CSR reports, which indicated to fulfil the requirements of those Performance Indicators, were manually placed under each category.

We carefully categorized the documents under their appropriate categories after analysing those CSR reports scored in year 2010, published on GRI website [65].

The Environmental corpus contained 593 sample documents altogether.

Performance Measure:

There are two types of results based on which we measured the performance of algorithms: the *atomic results* and the *aggregated* results.

Atomic result is the result of a single test executed on the dataset. Each execution records the following atomic results:

- True Positive (TP)
- False Positive (FP)
- False Negative (FN)
- Precision
- Recall

To minimize the negative effects of random selection and initialization approaches in some of the algorithms, each algorithm was executed 5 times; aggregated result is the arithmetic mean and standard deviation of the Recall rate of all executions.

Precision and Recall are two widely acceptable performance measures and we used them to make the effectiveness comparisons in our research. The following metric definitions are thus assumed:

$$Precision: P = \frac{TP}{TP + FP} \tag{1}$$

$$Recall: R = \frac{TP}{TP + FN} \tag{2}$$

The following aggregated metrics are assumed:

$$Effectiveness: E = \frac{1}{5}\sum_{n=1}^{5}(R_n) \tag{3}$$

Execution Platform:

We made use of Weka [18] and Rseslib [66] Java libraries to develop our custom desktop application using which we conducted the experiment. Since both libraries are also implemented on TunedIT.org [67] data mining platform, we made use of the platform's data mining features to conduct our experimental dataset analysis. The aggregated results of our experiments are publicly available on TunedIT.org Knowledge Base. These results are, therefore, fully reproducible.

Execution Results:

Fig. 1 shows the top-notch precision of Naïve Bayes, Neural Network and Random Subspace algorithms on our corpus while Decision Table algorithm delivered very fluctuating results with its precision on most categories falling far below 0.95. To measure the competency of our algorithms by tighter means, we decided to measure their Recall metric i.e. Their ability to correctly distinguish positive documents out of the entire testing set. As shown in Fig. 2, we witnessed a major drop of Random Subspace's Recall, but Naïve Bayes and Neural Networks managed to remain very effective.

It is clear that Neural Networks, Decision Table and Naïve Bayes have the highest effectiveness i.e. above 90% and Random Subspace produces the lowest effectiveness i.e. below 60%.

In order to show the difference between the produced Recall of algorithms in each execution, we calculated the Standard Deviation of those results. A bigger Standard Deviation means a less reliable algorithm as it has produced fluctuating Recalls. Fig. 3 illustrates a major Standard Deviation for Random Subspace, while Neural Networks, Decision Table and Naïve Bayes show Standard Deviations of near zero; this shows their almost identical performance of all iterations.

Fig. 4 draws the classification errors of all the classifiers at their best execution. The x axis represents the corpus, or expected classes and the y axis shows the actual classifier predictions. In this representation, a correct classification would draw a point p (x , y) with x = y. Any point outside the diagonal line i.e. with unequal x and y coordinates represents a classification error.

As shown in Fig.4, Naïve Bayes and Neural Networks classifiers produced the best results while Decision Tree and Random Subspace classifiers were less effective as they had more classification errors.

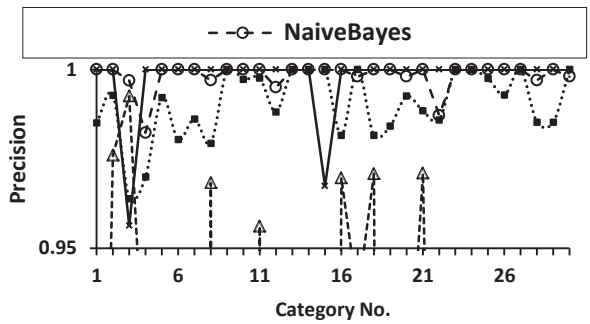

Fig.1. Precision of Algorithms on categories

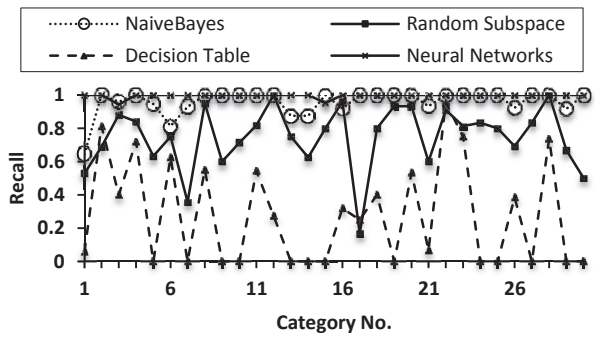

Fig. 2. Recall of Algorithms on Categories

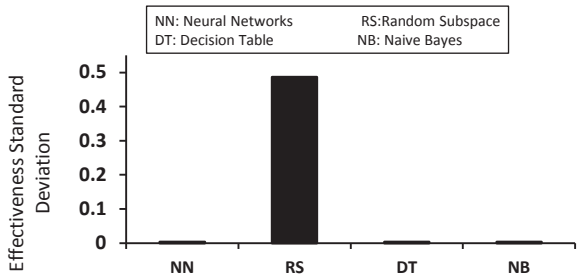

Fig. 3. Standard Deviation of Algorithms after 5 executions

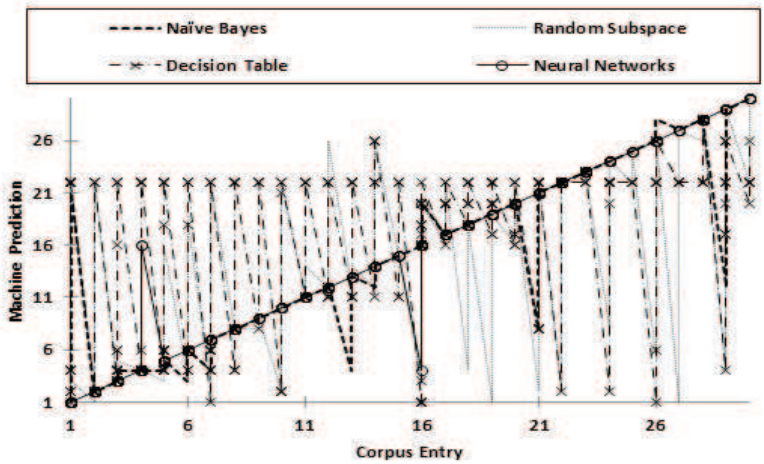

Fig. 4. Classification Errors

3.2. *Stage 2: Classification Enhancement*

Our findings of the previous stage made us favour Naïve Bayes classifier mainly due to its considerably higher learning speed compared to Neural Networks and its much greater effectiveness compared to the rest of the algorithms. We however witnessed that Naïve Bayes is prone to certain classification confusions among some categories. We also could not help but to notice that despite applying a stop-words pre-learning feature filter on our corpus, the size of the corpus had remained relatively large causing slower learning process, and perhaps confusion. In order to remedy these issues, we conducted two phases of enhancements i.e. 1) refine the document categories based on their respective documents' true ontological characteristics rather than on GRI's default categorization and 2) apply a heuristic feature selection algorithm to further reduce the size of the learning space.

3.2.1. *Training Categories Enhancement*

We performed a thorough ontological analysis on CSR report document performance indicators to discover those with conceptual similarities. Our studies showed that for instance, out of 30 performance indicators (categories) in the CSR Environmental

section, many share somewhat similar ontological characteristics. Therefore, we could further categorize those Performance Indicators under virtual super-categories (scopes), on which we built dedicated classifiers. We used Ontogen document ontology analysis software[68], to analyze our corpus documents. Table 1 shows the ontological characteristics of each scope along with the performance indicators which we placed under each scope due to their ontological similarities.

As a result, we constructed 4 (four) document classifiers: The first classifier was trained on a corpus with 5 categories i.e. scopes, which would determine the scope of each unlabelled document. The remaining 3 classifiers (called sub-classifiers) were individually trained on each scope, as a separate corpus, to learn to determine the exact performance indicator. Scopes 2 and 5 were excluded because of being unary i.e. with only one class. This made our application work as follows: After an unlabelled document's scope is determined by the first classifier, it is redirected to the corresponding sub-classifier to categorize the document under an appropriate performance indicator. This methodology helped us reduce the number of candidate categories for each document from 30 to maximum 21 categories and therefore improve the chance of a correct classification. Fig.5 below illustrates this workflow.

Four training corpuses were created in total; each for a classifier i.e. one for the scope finding classifier and one for each of the non-unary scopes. We carefully extracted sample documents from officially published corporate sustainability reports through more than 100 corporate websites as well as from the official CSR reports repository of Global Reporting Initiatives. In order to assure the quality of our corpus, we made sure to use the latest version of CSR reports which had received at least a B level score from either GRI or third party firms.

Table 1. Scopes and Performance Indicators along with their keywords

Scope	Performance Indicator	Keywords
1	EN11 to EN15	protected, biodiversity, impact, habitats, land, management, species, environmental, areas, companies
2	EN28	compliance, monetary, sanctions, environmental, related, regulations, laws, company, significant
3	EN3 to EN10, EN16 to EN27 and EN29	water, emissions, energy, waste, CO2, electricity, consumption, sources, discharges
4	EN1, EN2	paper, recycling, materials, consumption, waste, tons, total, raw
5	EN30	investments, protection, environment, million, expenditures

For corpus document categorization, we made use of the GRI content index attached to each CSR report.

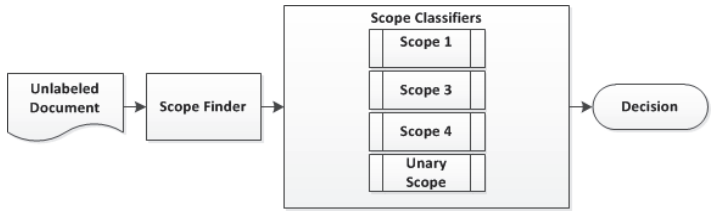

Fig. 5. Overall Workflow and Component Interaction Diagram

Our Corpuses contain more than 1000 sample documents altogether, which deemed to fulfil particular GRI performance indicator requirements. Table 2 shows the number of documents in each training corpus.

3.2.2. *Feature Selection Enhancement*

This time we used Correlation-based Feature Selection (CFS) algorithm[69] as our feature selection algorithm.

Similar to other heuristic feature selection algorithms, CFS performs its task by searching for good feature subsets and performing an evaluation to find the best subset. Among the most popular heuristic feature subset search algorithms are hill climbing and Best-First[70]. As Best-First has been proven to produce higher quality results[69], it remains as our chosen subset search method in this research.

The Best-First subset search algorithm starts by an empty feature set and generates a search tree of all possible single feature expansion subsets. The best evaluated feature subset is then selected as a candidate and the search continues to look for better candidates by expanding the subset in the same single expansion manner. If no improvement is observed, the search is taken to the next best candidate. The Best-First search algorithm will eventually return the best candidate subset after trotting through the entire search space, given sufficient time. See Fig.6 for an illustration of the internal structure of CFS algorithm and how it interacts with other components of the system.

Being a Correlation-based feature selection algorithm, CFS scores feature subsets based on their feature correlation to the class attribute and also to each other. It selects the best feature subset by giving high scores to subsets that contain features with high correlation to class attribute, but low correlation to each other. The following equation formalizes its heuristic:

$$G_s = \frac{K\,\overline{r_{ci}}}{\sqrt{k + k(k-1)\overline{r_{ii}}}} \tag{4}$$

Where: S denotes the subset to be merited, Kdenotes the number of its attributes, $\overline{r_{ci}}$ models the correlation of the attributes to the class attribute and $\overline{r_{ii}}$ denotes the inter-correlation between attributes.

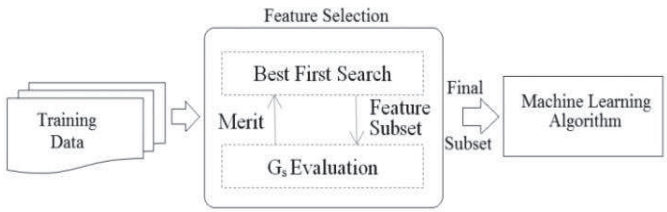

Fig. 6. Overall machine learning components interactions with CFS filter

Table 2. Number of positive sample documents per corpus

Training corpus	Number of positive samples
Scope Finder	1022
Scope 1	180
Scope 3	740
Scope 4	73

1.1.1 *Execution Results*

We took the train-and-test approach to test the document classification accuracy of our classifier; the training corpus was initially divided into two document sets for classification training and testing. The training and testing sets were initially populated with random documents from corpus, but we maintained the size ratio of both sets at 70% and 30% of total corpus size for training and testing sets respectively.

As can be seen in Table 3 below, the scope finder classifier has performed very well (around 92%) in determining the scope for novel documents using NaiveBayes algorithm after applying the CFS feature selection algorithm. The effectiveness of NaiveBayes has however dropped to about 55% in classification of scopes 1 and 3, but has regained its high accuracy on scope 4 to nearly 85%. Neural Networks, C45 and Decision Table algorithms were on the other hand proven to be much more effective than NaiveBayes when it came to selecting the exact performance indicators for documents. Fig. 7 and Fig. 8 illustrate the classification accuracy of our document classifiers after applying the stop words filter and CFS filter respectively.

The considerable drop of NaiveBayes efficiency on scopes 1, 3 and 4 is mainly due to high degree of ontological similarities between their underlying performance indicators. NaiveBayes was previously proven to suffer from confusion when learning the distinctive characteristics of ontologically similar categories. It, however, performs very well in classification of documents under ontologically distinctive categories such as CSR document scopes.

By observing the results, one could suggest the use of Neural Networks in determining the scope and performance indicators for arriving documents for optimal results, however the algorithm has shown to be extremely resource intensive and time consuming compared to others. As we prefer the software to function in a responsive manner, we would suggest to use NaiveBayes algorithm combined with CFS filter in

determining the scope of novel documents as the first classification step and performing the further performance indicator classification using either Neural Networks or Decision Table algorithms accompanied by stop words filter to produce highly accurate results.

Table 3. Effectiveness of Classifiers on Scopes Using Different Feature Selection Algorithms

Effectiveness / Algorithms			Classifier Effectiveness Rate (%)			
			Scope 4 Classifier	Scope 2 Classifier	Scope 1 Classifier	Scope Finder
Stop Words Filter	Classification Algorithm	NB[1]	71.69	52.86	47.36	87.62
		NN[2]	97.50	99.86	99.12	99.91
		C45	68.98	65.57	83.21	83.65
		DT[3]	96.25	79.29	96.49	82.89
CFS Filter	Classification Algorithm	NB	85.00	53.00	60	91.76
		NN	73.75	52.71	49.12	90.82
		C45	83.75	52.57	59.65	90.82
		DT	82.5	35.43	38.6	90.11

[1] NaiveBayes, [2] Neural Networks, [3] Decision Table

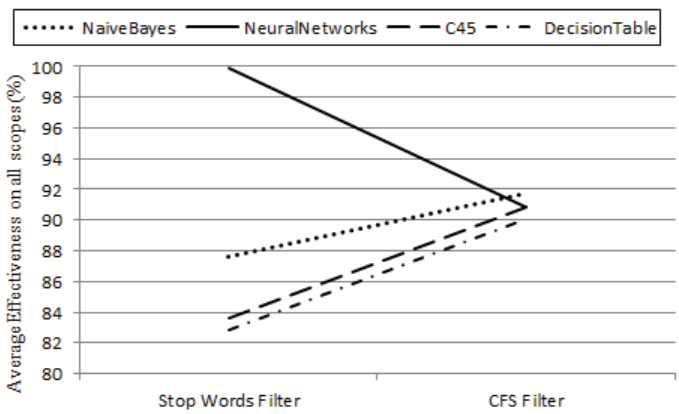

Fig. 7. Effectiveness of Scope Finder classifier using different feature filters

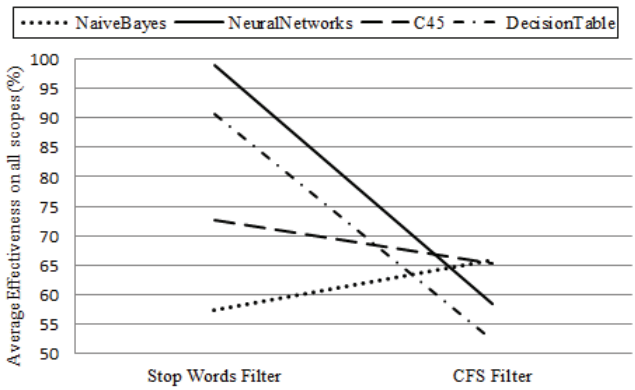

Fig. 8. Average effectiveness of scope classifier on all scopes using different feature filters

3.3. *Stage 3: Report Scoring*

Having completed the last two stages, we had chosen the most suitable classification and feature selection algorithms needed for implementation of the proposed CSR report scoring software. We designed the software to score the reports against the Global Reporting Initiative (GRI) version 3, also known as G3, framework. A brief discussion of the G3 scoring system follows.

3.3.1. *G3 Report Scoring System*

A GRI based corporate sustainability report shall report on at least 10 out of the 49 required disclosure items mandated by the framework in order to qualify for Application Level C. In order to qualify for higher levels of B or ultimately A, the report shall disclose at least 20 or all of the 49 mandatory disclosures respectively. The framework also contains some optional disclosure on which reporting entities might choose to report instead of the mandatory items if they aim at application levels B or C. Reporters may self-declare an application level based on the amount of disclosures in the report and publish through their preferred distribution channels. They could also go the extra mile and opt for having their application level claim get externally assured by GRI or a selected third-party firm authorized by GRI. Those reports which pass the external assurance step would be appended a "+" symbol i.e. their application level would be considered as C+, B+ or A+.

3.3.2. *Anatomy of the designed software*

The developed software solution to the problem of this research contains the following components: Firstly, A text processing module to manipulate, process and format the textual content of reports, secondly a text classification module to attempt to classify the text received from the text processor and lastly a report scoring module, which attempts to determine the application level of GRI framework to the processed text, based on the

input received from the classification module. See Fig. 9 and Fig. 12 below for a visual presentation of the main components of the solution, its sub-components and the process flow among them.

As can be seen if Fig.9, the Scoring Module contains two major components i.e. Scoring Framework class which contains an XML file describing the reporting framework used as well as fine details on its dimensions, sections and indicators and the Scoring Engine which compares a textual CSR report against the framework to determine the level to which the framework has been applied to the report. The Classification Module contains the manually tailored training corpus, the classifier classes(s) using which sections of the report are categorized under predetermined categories as well as the feature selection filter i.e. the Stop-Words filter and CFS filter. The text processing module handles text editing and manipulating operations which are mainly built-in Microsoft Word 2010 software. The Export Engine was developed and integrated into Word 2010 to integrate the support of saving the results of report scoring in Microsoft Excel 2010 file format.

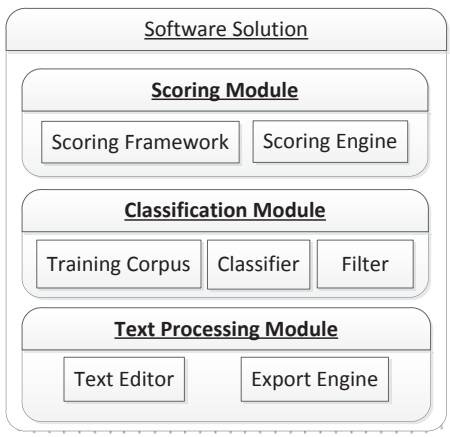

Fig. 9.Software components of the proposed GRI report scoring solution

Fig.12 draws the workflow of the internal components of the modules discussed and how they interact together from opening the CSR report to classifier construction and report scoring through performing the report scoring task and lastly exporting the results.

The software lets the user import a CSR report in PDF file format. CSR reports have no predefined format and structure therefore reporting entities have full flexibility on how, where and to what extent to disclose information. It is therefore safe to believe that the input to the software is completely unstructured when it comes to searching for a particular data (i.e. random access to data is not possible).

When a PDF report is imported, it is converted to Word format in an attempt to define a structure for it. This is done because, unlike PDF format, Word file format enjoys a hierarchical data structure. The elements of Word document structure from bottom to top are Range and Document respectively. The range object facilitates accessing document paragraphs, sentences, words and characters. This means possibility of accessing any

part of the document randomly needless of sequential search. For instance accessing the second sentence of the third paragraph page number 5 could be done randomly.

Importing a PDF document and converting it to Word format, therefore involves breaking the report into Pages, the pages are then broken down into paragraphs to be further broken down into sentence, words and finally characters. See Fig. 10 to grasp an idea on how a Word 2010 document is structured.

However, it is important to note that although the contents of the report are given a structure using the above methodology it will be still impossible to know the exact position of certain textual contents in reports mainly because, as mentioned earlier, CSR reports do not follow any standard content order and feature no content index. For example when looking for whether indicator 3 of the Environmental section (EN3) is disclosed, there would be no alternative to performing a blind paragraph-by-paragraph search starting from the first paragraph until a matching paragraph is found for it to be subject to automatic tagging using our developed Intelligent Tagging Engine (IntelliTag) or the search reaches the end of the document.

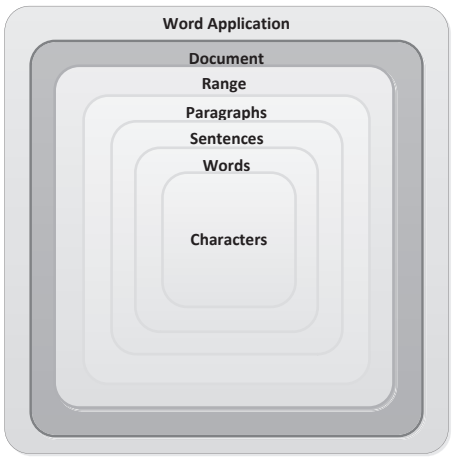

Fig. 10. Word Document Object Model

IntelliTag consists of three classifiers (i.e. one for each dimension) named as Economic Classifier, Environmental Classifier, and Social Classifier. These are all immediate children of the abstract Text Classifier Engine class. The class diagram in Fig. 11 presents this architectural idea visually.

Immediately after receiving user's command to commence the intelligent scoring process, IntelliTag starts to construct a classifier model using the supplied training corpus or de-serialises a pre-serialised model from the supplied binaries based on user preferences. After classifier model construction (or loading), IntelliTag iterates through the document on a paragraph by paragraph basis and treats each paragraph as a candidate document to be classified under either or none of the model categories (or disclosure items). This process is illustrated in Classification swim lane of Fig.12 below. In fact, a more detailed process flow diagram of the internal procedures of the three modules i.e.

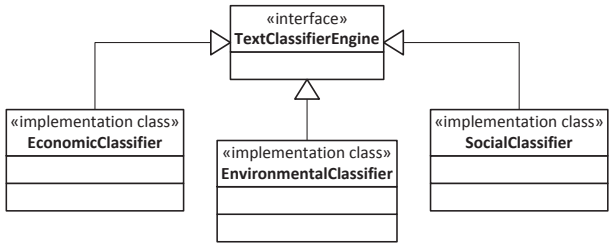

Fig. 11. Class Diagram of intelliTag

Scoring Module, Classification Module and the Text Processing Module are illustrated in Fig. 12 below.

3.3.3. *Effectiveness Measure*

As the aim of this research was to determine the effectiveness of our software system, we chose 100 externally assured and self-declared CSR reports to be scored by our software for their author-claimed scores (Ω) to be compared to scores determined by our software (Ω'). The following was assumed when conducting our tests:

Assumption 1: An automatic scoring would be successful if Ω' = Ω.

However, as our software is unable to perform an external assurance process on reports, it is not allowed to allocate a "+" symbol to the calculated application levels. It is

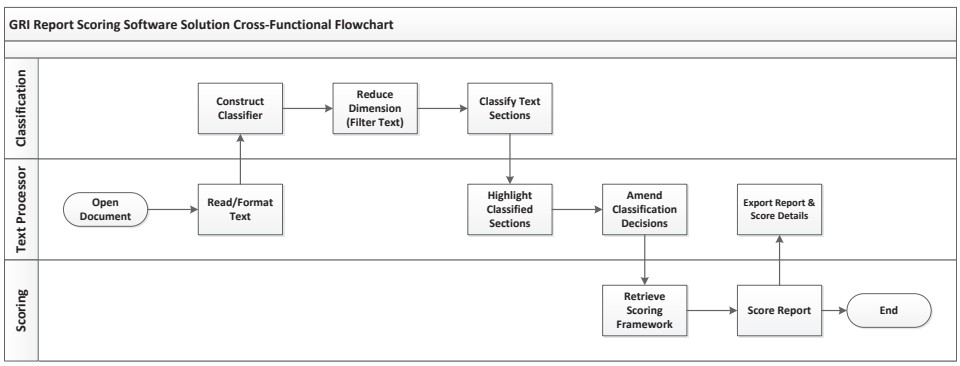

Fig. 12. Software Process Flow Model

therefore safe to ignore the "+" symbol when comparing the claimed and calculated scores. For instance, an automatic scoring of a report to application level B would be considered successful if the report's authors have published it with application level of B or B+.

Table 4 illustrates the frequency of claimed (or declared) application levels of reports we used for testing our software. Table 5 shows the tabulated data of Table 4 after assuming Hypothesis 1 above.

We would then apply the Pearson Correlation model to determine the correlation of scoring given by the organisations with software calculated scores.

In addition to testing the accuracy of overall application level calculation, we were also keen to know the effectiveness of our software in discovering information disclosures in the report. To do so, we picked 25 externally assured or GRI-checked reports which contained a GRI index of information disclosures and ran our software on them to figure out whether or not it is capable of discovering the disclosures mentioned in the report's attached GRI index table effectively. We used the following equation to calculate the Recall measure (R) of our software on each report as in equation (1)

The average software effectiveness was later calculated as in equation (3).

4. Experimental Results

As mentioned earlier, we measured the effectiveness of our software by comparing its accuracy in calculating the correct score (or application level) as well as discovering correct disclosure items in reports. These results are also referred to as *aggregated results* and *atomic results* respectively.

4.1. *Aggregated Results*

Aggregated results were calculated using Pearson product-moment correlation coefficient measure (Pearson *r*) by measuring the linear dependence between authors' claimed application levels (also known as *Equivalent application level*) and scores calculated by our software (also known as *WaveDive determined application level*).

Table 6 shows the results of running our software on the test reports. As can be seen there, none of the reports have received Application Level A while 42 of them were automatically score as Level B and 51 reports were determined to qualify for level C. It is also shown that the software has failed to determine the application level of 7 reports.

According to Table 5 below, 35 of the reports had been claimed to qualify for application level A and A+ by their authors; however, by looking into their attached GRI index tables we could not help but notice that none of them are inclusive of all the required disclosure items.

Table 4. Declared Application Levels of Tested Reports

Application Level	Frequency	Share (%)
A+	23	23.0
A+	12	12.0
B+	14	14.0
B+	24	24.0
C+	7	7.0
C+	20	20.0
Total	100	100.0

Instead, they contain references to external resources such as company websites for some disclosures. Although this method of reporting is totally accepted and permitted by GRI standards, the software is unable to follow the links to those external resources and fetch the resulting data for classification.

It means that according to the software, these reports do not contain sufficient disclosures -within them- to qualify for their declared application levels.

This limitation is the main cause of underscoring some of the tested reports to a lower application level. Nevertheless, no over-scoring (determining a higher application level than claimed) was witnessed.

Using Statistical Package for Social Sciences (SPSS) Pearson correlation (r) was calculated between WaveDive software determined application level and Equivalent application level. It is found that the correlation between the selected variables is 0.531 and is significant at 0.01. The correlation of 0.531 is considered 'moderate'.

4.2. *Atomic Results*

The atomic results were calculated by comparing the discovery of information disclosure by software and actual disclosure claims by reporting entities on a disclosure-by-disclosure basis. A disclosure discovered by software which has been claimed to be reported by report author is counted as a True Positive while skipping a claimed disclosure is counted towards False Negatives. The Recall measure for each report dimension is calculated as in equation (1) before calculation of overall accuracy as in equation (2).

Table 7 below shows the number of true positives, wrong negatives and Recall measure of the software in discovering disclosed information in each dimension of the selected 25 reports. These results are illustrated visually in Fig. 13 below.

Table 5. Equivalent Application Levels of Tested Reports

Application Level (Ω)	Frequency	Share (%)
A	35	35.0
B	38	38.0
C	27	27.0
Total	100	100.0

Table 6. WaveDive Calculated Application Levels

Application Level	Frequency	Percentage	Valid Percentage	Cumulative Percentage
B	42	42.0	42.0	42.0
C	51	51.0	51.0	93.0
None	7	7.0	7.0	100.0
Total	100	100.0	100.0	

As expected, effectiveness on Environmental dimension was higher than on other dimensions mainly due to adapting a chain classification approach and combination of stop-words and correlation based feature selection.

The Social classifier is second most effective followed by Economic classifier, which shows an overall effectiveness of 73.71%. This happened despite higher number of Economic training samples compared with those for the Social classifier. We believe that it is due to high ontological similarities between Economic disclosures, which cause classification confusion.

The Social dimension, although bigger in terms of number of categories, has several ontologically-distinctive category clusters. This enables composing a more efficient classifier model and producing better results.

It is wise to calculate the overall effectiveness of the software in information discovery as an average of Recall measures in Table 7.

Therefore, information discovery effectiveness (E) is:

$$E = 81.10\ \%$$

Apart from testing the effectiveness of the algorithms, we also conducted modular unit testing on all on class objects of the system by providing each module with sample inputs and comparing their produced results of execution with expected outcomes. The reliability of system modules were tested by a brute force data injection and observing their reaction as well as the produced results. This method of testing ensures that the system is stable when facing unforeseen exceptions and produces reliable results if provided with healthy input.

Table 7. Atomic Results of Execution on selected reports

Dimension	No. of indicators	N	TP[1]	FN[2]	Recall (%)
Economic	7	25	129	46	73.71
Environmental	17	25	391	34	92.00
Social	25	25	485	140	77.60

1: No. of True Positives

2: No. of False Negatives

5. Statistical Analysis of Discovered Disclosures

In addition to streamlining the process of analyzing and scoring CSR reports, WaveDive software might as well be used for performing various statistical studies on different dimensions of CSR reporting such as discovering reporting behaviors, habits and patterns across organizations. Although the number of reports on which this study was undertaken was relatively small (i.e. 100 reports) and therefore a solid conclusion could not be drawn regarding reporting practices of organizations, this section attempts to shed a light on the possibilities having this kind of data creates to open the way for future research.

5.1. *Materials and Methods*

In order to facilitate such studies, WaveDive supports exporting the performance indicators discovered (and those tagged manually) to an Excel 2010 workbook. The workbook contains two (2) worksheets which contain a full list of exported disclosure items and certain extra information regarding the report itself. This extra information is

obtained from user upon exporting to Excel using the designated Extra Information Windows Form. Table 8 and Table 9 below visualize the structure of each of the worksheets.

Having executed WaveDive software on 100 CSR test reports and exported the results of them to Excel format, 100 Excel workbooks each of which containing 2 worksheets were created.

Table 8. Data Structure of Worksheet No.1

Field Name	Data Type/Possible Entries	Remarks
Performance Indicator	String (4 characters) e.g. EN01	Initials of performance indicators
Description	String (Free format)	Long description of performance indicator according to framework
Reported	Boolean (Yes/No)	Indicates whether the indicator is discovered or manually tagged
Cross Reference	String (Free format)	Page number(s)

In order to perform the analysis in a quicker pace, the workbooks were merged into a single Excel workbook. The new workbook, as a result, included 200 worksheets (2 for each report). This new Excel workbook is referred to as the Facts Workbook from now on.

Table 9. Data Structure of Worksheet No.2

Field Name	Data Type/Possible Entries
Nominated Application level	A+, A, B+, B, C+, C
Determined Application level	A, B, C, None
Status	GRI Checked, Third-Party Checked, Not Checked
Company Size	Large, MNE, SME
Listed Company	Yes, No
Organization Type	Non-Profit, Partnership, Private, Public, State-Owned, Subsidiary
Sector	Production, Service, Trade, Other
Supplementary Kit	Not Applicable, Not Used, Used
Region	Africa, Asia, Europe, Latin America, Northern America, Oceania
Number of Unique Performance Indicators Discovered (NUPI)	Integer value e.g. 7
Number of Required Disclosures for Selected Application Level (NRDAL)	Integer value e.g. 20
Total Number of Required Disclosure Items (TNRDI)	Integer value e.g. 49 if using GRI G3

5.2. *Analysis Results*

Below comes a series of findings made as a result of conducting the statistical studies on the Facts Workbook.

5.2.1. *Performance Indicator Popularity*

The data gathered in the Facts Workbook shows that EC1 is the most popular indicator on which 91 out of the 100 reports have reported. Performance indicator LA2 on the other hand is the least popular as no corporation had been found to have reported on it.

Performance Indicator EC1 is described by G3 framework as ”Direct economic value generated and distributed, including revenues, operating costs, employee compensation, donations and other community investments, retained earnings, and payments to capital providers and governments”.

Performance Indicator LA2 is described as “Total number and rate of employee turnover by age group, gender, and region”.

Section Popularity across Industries:

Another interesting study is to figure out the popularity of performance indicator sections (i.e. Economic, Environment, and Society) across companies of various industries. As for the data gathered in this research, it was found that Production sector companies tend to report more on their environmental performance while service companies report more on their economic performance in regards to sustainable development. See Fig.13 below.

Correlation between report information variables:

Yet another interesting study on the gathered data would be to unleash the relations between pairs of report variables (also known as variable correlation). The results of such study reveal the significance of variables’ influence on one another. For this reason, the Parson Correlation and Chi-Square analysis were done on the variables. The results of this study are presented in Table 10 below. Refer to Table 9 for acronyms.

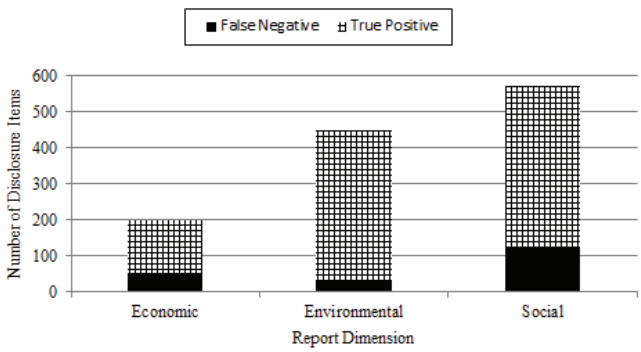

Fig. 13. Atomic Results of Execution

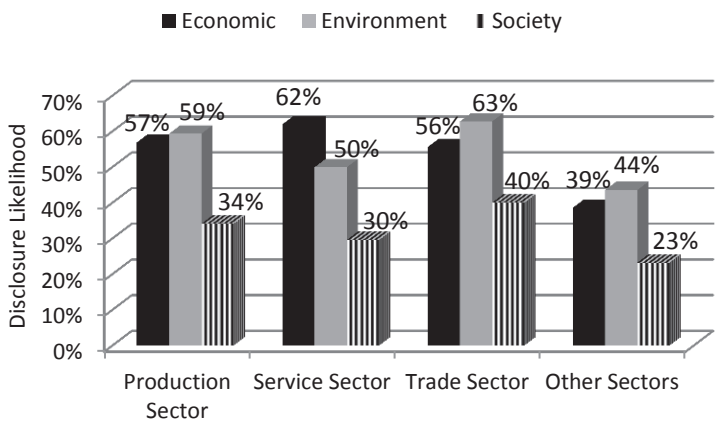

Fig. 14. Popularity of different CSR report sections among various industries

Among the most interesting findings are the moderately significant correlation between nominated application levels and the number of pages per report, the slightly significant correlation between the listed status of companies with the number of discovered disclosures, and the high influence of business sector and regions on nominated application levels.

Table 10. Pearson Correlations and Chi-Square (In Brackets) For the Selected Variables

	NUPI	NRDAL	NUPI/ NRDAL	NUPI/ TNRDI	No. of pages
Claimed Application level	0.117 *(-0.536)*	0.000 *(-.933)*	0.000 *(0.705)*	0.117 *(-.536)*	0.088 *(-.133)*
Determined Application Level	0.000 *(-.676)*	0.000 *(-0.478)*	0.135 *(.102)*	0.000 *(-0.676)*	0.305 *(-.271)*
Report External Assurance Status	0.357 *(-.254)*	0.122 *(-.233)*	0.359 *(.088)*	0.357 *(-.254)*	0.295 *(0.007)*
Company Size	0.495 *(-0.001)*	0.254 *(0.007)*	0.713 *(0.048)*	0.495 *(-0.001)*	0.316 *(-0.156)*
Listed Company	0.096 *(-0.262)*	0.985 *(-0.012)*	0.378 *(-0.169)*	0.096 *(-0.262)*	0.288 *(-0.068)*
Organization Type	0.923 *(.203)*	0.469 *(0.207)*	0.211 *(-.118)*	0.923 *(0.203)*	0.893 *(0.113)*
Sector	0.165 *(-.202)*	0.001 *(-0.188)*	0.122 *(.118)*	0.165 *(-0.202)*	0.192 *(-0.081)*
Supplements	0.813 *(0.082)*	0.352 *(0.155)*	0.622 *(-0.127)*	0.813 *(0.082)*	0.365 *(0.105)*
Region	0.415 *(-0.109)*	0.02 *(-.136)*	0.582 *(0.117)*	0.415 *(-.109)*	0.511 *(-0.131)*

It can be also observed that the type and size or organizations as well as whether or not they had applied GRI supplementary kits in their reporting have little or no effect on other variables.

6. Conclusion

This article demonstrates the details of our research into developing an automated solution for corporate sustainability report scoring. After looking into the state of the art in artificial intelligence as well as studying similar problems and solutions we picked machine learning approach to text categorization as the solution to tacking the longstanding problem.

The research continued by finding the most suitable classification and feature selection algorithms. Powered by the most suitable supervised machine learning algorithms and a training corpus containing thousands of sample disclosures, the software managed to yield a considerably high accuracy rate in discovering disclosure items in reports at 81.10%. Furthermore, the results of running it on the test set reports showed that the software generates moderately accurate results when it comes to determining application levels for the reports despite certain limitations and constraints handling many of which were outside the scope of the project.

In order to show that the usefulness of the software is not limited only to scoring CSR reports, a handful of statistical analysis studies were conducted on the results produced by the software which unleashed interesting findings regarding the tested reports such as popularity of certain performance indicators as well as volume of disclosures made by reporting companies across different business sectors. Other analysis studies conducted in this research include discovering the average number of full disclosures in each report section which highlighted an obvious bias by reporting organizations towards reporting more on environmental and economic aspects than on social as well as a brief study of correlations between various report variables such as the nominated and determined application levels, company sizes, industries, size of reports, etc.

All in all, the findings discussed in this article show suitability of the undertaken approach to CSR report analysis and also sheds light on unlimited research possibilities in brings along, be it technical or analytical.

References

1. A. Nutz and M. Strauss, "Concept and practical implementation of the eXtensible Business Reporting Language (XBRL)," Wirtschaftsinformatik, vol. 44, pp. 447-457, Oct 2002.
2. H.-K. Arndt, R. Isenmann, J. Brosowski, I. Thiessen, and J. Marx-Gomez, "Sustainability Reporting Using the eXtensible Business Reporting Language (XBRL)," Managing Environmental Knowledge, pp. 75-82, 2006.
3. T. Suzuki, "XBRL processor "interstage Xwand" and its application programs," Fujitsu Scientific & Technical Journal, vol. 40, pp. 74-79, 2004.
4. S. Briciu, L. S. Todor, and H. T. Andreica, "Xbrl - an Efficient Method of the Accounting Reporting Management," Metalurgia International, vol. 15, pp. 125-128, 2010.
5. S. Leibs. (2007). Sustainability Reporting: Earth in the Balance Sheet [Article]. Available: http://www.feicanada.org/files/Environmental%20Sustainability%20article.doc

6. R. Thurm, "Taking the GRI to Sclae," in Sustainability Accounting and Reporting, ed: Springer, 2006, pp. 325-337.
7. F. Sebastiani, "Machine learning in automated text categorization," ACM Comput. Surv., vol. 34, pp. 1-47, 2002.
8. J. Weizenbaum, "Contextual understanding by computers," Communications of the Acm, vol. 10, pp. 474-480, 1967.
9. C. C. Shilakes and J. Tylman, "Enterprise information portals," Merrill Lynch, November, vol. 16, 1998.
10. S. Chakrabarti, B. Dom, R. Agrawal, and P. Raghavan, "Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies," Vldb Journal, vol. 7, pp. 163-178, Aug 1998.
11. J. Hroza, J. Zizka, and A. Bourek, "Filtering very similar text documents: A case study," Computational Linguistics and Intelligent Text Processing, vol. 2945, pp. 511-520, 2004.
12. R. L. Liu, "Dynamic category profiling for text filtering and classification," Information Processing & Management, vol. 43, pp. 154-168, Jan 2007.
13. C. J. Fall, A. Torcsvari, P. Fievet, and G. Karetka, "Automated categorization of German-language patent documents," Expert Systems with Applications, vol. 26, pp. 269-277, Feb 2004.
14. H. Mathiassen and D. Ortiz-Arroyo, "Automatic categorization of patent applications using classifier combinations," Intelligent Data Engineering and Automated Learning - Ideal 2006, Proceedings, vol. 4224, pp. 1039-1047, 2006.
15. J. Abernethy, O. Chapelle, and C. Castillo, "Graph regularization methods for Web spam detection," Machine Learning, vol. 81, pp. 207-225, Nov 2010.
16. P. Cortez, C. Lopes, P. Sousa, M. Rocha, and M. Rio, "Symbiotic filtering for spam email detection," Expert Systems with Applications, vol. 38, pp. 9365-9372, Aug 2011.
17. A. Lad, "SpamNet - Spam detection using PCA and neural networks," Intelligent Information Technology, Proceedings, vol. 3356, pp. 205-213, 2004.
18. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," SIGKDD Explorations, vol. 11, 2009.
19. J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang, "Predicting protein–protein interactions based only on sequences information," Proceedings of the National Academy of Sciences, vol. 104, pp. 4337-4341, 2007.
20. S. Jarvis, "Data mining with learner corpora," A Taste for Corpora: In Honour of Sylviane Granger, vol. 45, 2011.
21. A. Wilson and P. Rayson, "Automatic content analysis of spoken discourse: a report on work in progress," Corpus based computational linguistics, pp. 215-226, 1993.
22. P. J. Stone, D. C. Dunphy, and M. S. Smith, "The General Inquirer: A Computer Approach to Content Analysis," 1966.
23. R. Tesch, Qualitative research: Analysis types and software tools: Routledge, 1990.
24. S. A. Crossley and D. S. McNamara, "Computational assessment of lexical differences in L1 and L2 writing," Journal of Second Language Writing, vol. 18, pp. 119-135, 2009.
25. V. Koller, A. Hardie, P. Rayson, and E. Semino, "Using a semantic annotation tool for the analysis of metaphor in discourse," Metaphorik. de, vol. 15, pp. 141-160, 2008.
26. P. Rayson, "New trends in corpus linguistics for translation studies," in CCID & Lancaster University Workshop on Corpus Linguistics & Machine Translation Applications, 2008.
27. O. Mudraya, B. Babych, S. Piao, P. Rayson, and A. Wilson, "Developing a Russian semantic tagger for automatic semantic annotation," Corpus Linguistics 2006, pp. 290-297, 2006.
28. M. Deegan, H. Short, D. Archer, P. Baker, T. McEnery, and P. Rayson, "Computational Linguistics meets Metadata, or the automatic extraction of key words from full text content," RLG Diginews, vol. 8, 2004.
29. D. McDonald, I. McNicoll, G. Weir, T. Reimer, J. Redfearn, N. Jacobs, and R. Bruce, "Value and benefits of text mining," JISC Digital infrastructure (2012), 2012.

30. P. J. Herron, "Text Mining Adoption for Pharmacogenomics-based Drug Discovery in a Large Pharmaceutical Company: a Case Study," Unpublished master's thesis, University of North Carolina, Chapel Hill, 2006.
31. H. Jung, P. Kim, S. Lee, M. Lee, D. Seo, and W. Sung, "iLaw system: intelligent legislation support system based on semantic web and text mining technologies," in Proc. 5th International Conference on Methodologies, Technologies and Tools enabling e-Government (MeTTeG 2011), Italy, 2011, pp. 1-12.
32. P. Thompson, "Text mining, names and security," Journal of Database Management (JDM), vol. 16, pp. 54-59, 2005.
33. A. Kolk, "Trends in sustainability reporting by the Fortune Global 250," Business Strategy and the Environment, vol. 12, pp. 279-291, 2003.
34. Global-Reporting-Initiative. (2010, 19/09/2011). GRI Reporting Statistics. Available: http://www.globalreporting.org/NR/rdonlyres/954C01F1-9439-468F-B8C2-B85F67560FA1/0/GRIReportingStats.pdf
35. L. S. Mahoney, L. Thorne, L. Cecil, and W. Lagore, "A research note on standalone corporate social responsibility reports: Signaling or greenwashing?," Critical Perspectives on Accounting, 2012.
36. Y. Sumiani, Y. Haslinda, and G. Lehman, "Environmental reporting in a developing country: a case study on status and implementation in Malaysia," Journal of Cleaner Production, vol. 15, pp. 895-901, 2007.
37. D. Wheeler and J. Elkington, "The end of the corporate environmental report? Or the advent of cybernetic sustainability reporting and communication," Business Strategy and the Environment, vol. 10, pp. 1-14, 2001.
38. R. Scott Marshall and D. Brown, "Corporate environmental reporting: what's in a metric?," Business Strategy and the Environment, vol. 12, pp. 87-106, 2003.
39. G. Azzone, M. Brophy, G. Noci, R. Welford, and W. Young, "A stakeholders' view of environmental reporting," Long Range Planning, vol. 30, pp. 699-709, 1997.
40. P. Cerin, "Communication in corporate environmental reports," Corporate Social Responsibility and Environmental Management, vol. 9, pp. 46-65, 2002.
41. P. Rao, A. K. Singh, O. la O'Castillo, P. S. Intal Jr, and A. Sajid, "A metric for corporate environmental indicators… for small and medium enterprises in the Philippines," Business Strategy and the Environment, vol. 18, pp. 14-31, 2009.
42. D. M. Hussey, P. L. Kirsop, and R. E. Meissen, "Global reporting initiative guidelines: an evaluation of sustainable development metrics for industry," Environmental Quality Management, vol. 11, pp. 1-20, 2001.
43. L. Dgilienė and R. Gokienė, "Valuation of corporate social responsibility reports," Economics and management, vol. 2011, pp. 21-27, 2011.
44. U. Rosenström and J. Lyytimäki, "The role of indicators in improving timeliness of international environmental reports," European Environment, vol. 16, pp. 32-44, 2006.
45. J. E. Morhardt, S. Baird, and K. Freeman, "Scoring corporate environmental and sustainability reports using GRI 2000, ISO 14031 and other criteria," Corporate Social Responsibility and Environmental Management, vol. 9, pp. 215-233, 2002.
46. Global-Reporting-Initiative. (2011, 11/07/2011). What is GRI. Available: http://www.globalreporting.org/AboutGRI/WhatIsGRI/
47. R. Isenmann and C. Lenz, "Customized corporate environmental reporting by internet-based push and pull technologies," Eco Management and Auditing, vol. 8, pp. 100-110, 2001.
48. J. R. Modapothala and B. Issac, "Study of economic, environmental and social factors in sustainability reports using text mining and Bayesian analysis," in Industrial Electronics & Applications, 2009. ISIEA 2009. IEEE Symposium on, 2009, pp. 209-214.
49. G. Harte, L. Lewis, and D. Owen, "Ethical investment and the corporate reporting function," Critical Perspectives on Accounting, vol. 2, pp. 227-254, 1991.

50. J. R. Modapothala and B. Issac, "Analysis of corporate environmental reports using statistical techniques and data mining," 2009.
51. J. Modapothala, B. Issac, and E. Jayamani, "Appraising the Corporate Sustainability Reports–Text Mining and Multi-Discriminatory Analysis," Innovations in Computing Sciences and Software Engineering, pp. 489-494, 2010.
52. J. R. Modapothala and B. Issac, "Evaluation of corporate environmental reports using data mining approach," in Computer Engineering and Technology, 2009. ICCET'09. International Conference on, 2009, pp. 543-547.
53. T. Botsis, T. Buttolph, M. D. Nguyen, S. Winiecki, E. J. Woo, and R. Ball, "Vaccine adverse event text mining system for extracting features from vaccine safety reports," Journal of the American Medical Informatics Association, 2012.
54. M. Eckstein, "Working Title: Text Mining for a Database Containing Structured Reports about Nosocomial Outbreaks (Outbreak Database)," 2010.
55. C. Y. Shirata, H. Takeuchi, S. Ogino, and H. Watanabe, "Extracting Key Phrases as Predictors of Corporate Bankruptcy: Empirical Analysis of Annual Reports by Text Mining," Journal of Emerging Technologies in Accounting, vol. 8, pp. 31-44, 2011.
56. A. K. Prasad, S. Ramakrishna, D. S. Kumar, and B. P. Rani, "Extraction of Radiology Reports using Text mining," International Journal, vol. 2, 2010.
57. J. Friedlin, M. Mahoui, J. Jones, and P. Jamieson, "Knowledge Discovery and Data Mining of Free Text Radiology Reports," in Healthcare Informatics, Imaging and Systems Biology (HISB), 2011 First IEEE International Conference on, 2011, pp. 89-96.
58. R. Jensen and Q. Shen, Eds., Rough and Fuzzy Approaches (Computational Intelligence and Feature Selection. Hoboken, New Jersey: IEEE Press, 2008, p.^pp. Pages.
59. X. P. Li, K. Q. Shen, C. J. Ong, Z. Hui, and E. P. V. Wilder-Sniith, "A feature selection method for multilevel mental fatigue EEG classification," Ieee Transactions on Biomedical Engineering, vol. 54, pp. 1231-1237, Jul 2007.
60. C. Leilei, G. Hui, and C. Wenbo, "A new feature weighting method based on probability distribution in imbalanced text classification," in Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on, 2010, pp. 2335-2339.
61. Z. X. Li, Z. Y. Xiong, Y. F. Zhang, C. Y. Liu, and K. A. Li, "Fast text categorization using concise semantic analysis," Pattern Recognition Letters, vol. 32, pp. 441-448, Feb 1 2011.
62. R. Kohavi, "The power of decision tables," Machine Learning: Ecml-95, vol. 912, pp. 174-189, 1995.
63. T. K. Ho, "The random subspace method for constructing decision forests," Ieee Transactions on Pattern Analysis and Machine Intelligence, vol. 20, pp. 832-844, Aug 1998.
64. W. Li, B. Lee, F. Krausz, and K. Sahin, "Text Classification by a Neural Network," in Proceedings of the 1991 Summer Computer Simulation Conference. Twenty-Third Annual Summer Computer Simulation Conference, 1991, pp. 313-318.
65. Global-Reporting-Initiative. (2011, 19/09/2011). GRI Reports List. Available: http://www.globalreporting.org/ReportServices/GRIReportsList/
66. J. G. Bazan and M. S. Szczuka, "RSES and RSESlib - A Collection of Tools for Rough Set Computations," presented at the Revised Papers from the Second International Conference on Rough Sets and Current Trends in Computing, 2001.
67. M. Wojnarski, S. Stawicki, and P. Wojnarowski, "TunedIT.org: System for Automated Evaluation of Algorithms in Repeatable Experiments," in Rough Sets and Current Trends in Computing (RSCTC). vol. 6086, ed: Springer, 2010, pp. 20-29.
68. B. Fortuna, M. Grobelnik, and D. Mladenic, "OntoGen: Semi-automatic Ontology Editor," in HCI International 2007, Beijing, China, 2007.
69. M. A. Hall and L. A. Smith, Practical feature subset selection for machine learning vol. 20. Singapore: Springer-Verlag Singapore Pte Ltd, 1998.
70. E. Rich and K. Knight, Artificial Intelligence: McGraw-Hill, 1994.