An Insight into Corporate Social Responsibility Reports: A Text Mining Approach

Sourav Das (cs1140258@cse.iitd.ac.in)
Indian Institute of Technology Delhi, India

Alok Choudhary
Loughborough University United Kingdom

Jenny Harding
Loughborough University United Kingdom

Summary Abstract

Corporate Social Responsibility (CSR) report contains secondary data about company's intentions, strategies and activities along with their impact in supply chain management which is of great interest to stakeholders. Studies shows that a CSR report contains more than 50 papers on an average which is a large data to be analyzed and compared manually. This study focuses on analyzing the CSR reports from automotive industries using Multinomial Naïve Bayes algorithm from text mining. This study is on extracting concise data from large CSR reports from various companies to analyse the general trend over the years and to compare them according to year, revenue size, geographical location, political situations and market structure and also to study their impact on triple bottom line. It also includes finding relation between various sustainability areas including their relative strength using probabilistic text mining approach.

Keywords: Sustainability, Text mining, Automotive industry, Corporate social responsibility reports

INTRODUCTION: These days sustainability has become the key agenda of every industry. Most used definition of sustainable development says: It is the development that meets the needs of present without compromising the ability of future generations to meet their own needs (Our common future, Oxford 1987). By engaging in sustainability, companies can improve their image and generate better stakeholder attitude and support behavior, including enhanced stakeholders advocacy behaviors (Du et al., 2010). Moreover stakeholder demands should be seen as opportunities rather than constraints (Kurucz et al., 2008). More and more companies are disclosing details of their environmental performance in response to stakeholder demands in documents published on the internet, called Corporate Social Responsibility (CSR) or sustainability reports (Jose & Lee, 2007). CSR report contains secondary data about company's intentions, strategies, activities along with their impact in supply chain management which is of great interest to stakeholders (Tate et al., 2010). CSR has been acknowledged as one of the most important factors in determining corporate reputation (Worcester, 2007). Eager interest of stakeholders in engagement of companies in CSR is creating difficulties for researcher and practitioners to understand the specific consequences of CSR activities(Vlachos et al., 2009). Studies shows that average number of pages in a CSR report is more than 50 (Trembly et al., 2016) which is too much data to analyse manually and compare.

This study expands the research in three ways, first from literature review we devised 10 major areas about which CSR reports of automotive industries are talking about. We use text mining techniques to automatically classify content of large CSR reports into various categories (although similar techniques could be used to analyse CSR reports from any sector). Once categorized automatically with more than 75% of accuracy we use the data to look at the trend of sustainability among the industries. Secondly, this study will provide a quantitative measure to compare CSR trends among various industries depending upon year, geographical location, size, market structure etc. Finally it also gives relationship pattern between various sustainability areas with their relation strength in a quantitative way, which we have calculated using probabilistic approach. This data will be of great importance to companies to balance their limited resources in the sustainability areas to optimize their profit along with reputation in the market. The reason behind choosing automotive industry is because, it is the largest manufacturing enterprise in the world and one of the most resource intensive industries among all major industrial system (Amrina & Yusof, 2011).

This paper is organized as follows. The next section describes the current state of the art followed by describing our methodology, results and analysis, discussion and finally, limitations and future scope of this study.

LITERATURE REVIEW: Studies show that in every sector regardless of the industry, CSR reports focus on meeting stakeholders demand (Tate et al., 2010). So it is important for the stakeholders to know and understand the Triple Bottom Line (TBL) concerns of the industry. With data predicted to grow by 40% p.a., computational based data analysis has a significant potential in societal and economic improvement in the coming years and it has already been applied to many areas like biomedical sciences, chemistry and some early adoptions in humanities and social sciences (McDonald et al., 2012). As the number of CSR reports are increasing this is also an interesting and important area in which text mining can be applied. Global Reporting Initiative (GRI) reporting statistics show that 1800+ companies have produced and published CSR reports in 2010 which includes European Union’s 500 Global companies also (GRI 2011). Another survey also shows that 95% of world’s largest companies produce CSR reports (Mahoney et al., 2013). Studies also show that pressure from local governments and legislation also has a key influence in the decisions regarding CSR (Kolk, 2003). Application of text mining in analyzing CSR reports is new. Some recent studies has been found understanding sustainability using text mining (Modapothala et al., 2009; Wan et al., 2014; Rivera et al., 2014; Chae, 2015; Nishant et al., 2015). Modapothala et al. (2009) had applied Bayesian and text mining to score corporate environmental reports in terms of economic, environmental and social performance indicators using GRI guidelines. Wan et al. (2014) has applied text mining in process industries to understand sustainability trends. He studied four major process industries namely oil/petrochemicals, bulk/specialty chemicals, pharmaceuticals and consumer products. Rivera et al. (2014) applied supervised learning algorithm to track sustainability indicators by analyzing unstructured news articles. Chae, (2015) applied text mining in social network, twitter data to analyze supply chain tweets, highlighting the potential role of twitter in supply chain practice and research. Nishant et al, (2015) applied Centering Resonance Analysis (CRA) to analyze sustainability trend in Indian firms. But very limited study has been carried out in understanding CSR reports via text mining (Soiraya, 2011; Saha et al., 2015; Tremblay et al., 2016). We also could not find any methodology to compare the CSR concerns of any companies in any sector which is a huge research gap. Soiraya, (2011) applied text mining in CSR reports of 50 largest (by revenue 2007-11) IT industries to create a ontology from the content of CSR reports. Saha & Nabareseh, (2015) used text mining on CSR reports of Banks from India and Ghana to analyze company’s behavior towards CSR based on their disclosure practices. Wati & Koo, (2010) manually analyzed environmental reports of 4 big IT companies to draw concepts of green IT. Tate et al. (2010) applied Centering Resonance Analysis (CRA) for content analysis of 100 companies CSR reports. Shahi et al. (2014) used supervised learning algorithms on CSR reports to score the reports and compare them with the scores given by the authors. Tremblay et al., (2016) used both supervised and unsupervised learning algorithm on CSR reports to find evidence of importance of environmental concern in corporate policy. Panayiotou et al. (2009) analyzed 28 CSR reports published in Greece manually to find the different areas a CSR report focuses. In our study, the sustainability key performance areas are initially categorized with the help of literature review. A training set will be used to train the program to split articles into various categories. After training, the program will be able to identify the areas in which a CSR report focuses. The automotive industry web sites also showed recognition of increase in social expectations, projection of a desired corporate identity which makes it an important area to work on (Rolland et al, 2010). The reason behind choosing only automotive industries is because of subject expertise, uniformity in categorization and also because this area is till now untouched by researchers.

METHODOLOGY: Whole methodology is divided into of 5 parts, categorization, preprocessing, computerized classification, manual evaluation and relation extraction.

Categorization is completely based on key performance indicators, expert advice and content analysis studies of CSR reports from automotive sector as presented by various studies (Amrina & Yusof, 2011; Sukitsch et al., 2015).Sustainability practices are industry specific as different industries will face different challenges and will have different goals. For example, major sustainability issues faced by electronic industry are human rights labor violations and product recycling, whereas automotive industry's main concern will be fuel, energy efficiency and emission control (Accentre CEO study, 2011). Studies show that for an industry, implementation of sustainability leads to improvements of triple bottom line i.e. economy, social and environmental responsibility (Seuring & Muller, 2008). Careful measures have been taken while considering the specifications for each category.

Economic	**1.**	**4 R's (Reuse, Reduce, Recycle, Remanufacture)** • Reduction of raw materials • Reduction of labor forces • Time saving in manufacturing • Life cycle assessment issues • Increase in recyclability and reducing waste • Reducing harmful chemical use
	2.	**Business Development, Supply Chain Management** • Sustainability focus on dealers and suppliers • Packaging and transportation issues • New introduction of business plans • Flexibility of the industry • Improvement in logistics
	3.	**R&D, Energy and Fuel Efficiency** • Innovations • Efficiency in energy utilizations • Increase in fuel efficiency of vehicle • Technological advancement in vehicles
Environmental	**4.**	**Waste Disposal** • Waste treatment • Advancement in disposal techniques
	5.	**Emissions Control (Land, Water, Air)** • Environmental activities like reforestation, nature walk • Creating awareness among employees and consumers • Award and position in sustainability areas • Investigation on sustainable impacts • Creation of special committees concerning environment
	6.	**Noise Control** • Noise Control in industries • Reduction in noise in manufactured products • Reduction in noise in surrounding areas
	7.	**Community Involvement and Award** • Organisation of various environmental activities • Creating awareness among employees and consumers • Awards and positions in various areas • Investigations on various areas • Formation of committees
	8.	**Green Building & Green Manufacturing** • Green building and green manufacturing • Construction of new eco-infrastructure • Overall environmental concern

Social	**9.**	**Non-Employee Concern** • Consumer satisfaction • Government and NGO satisfaction • Safety and comfort of customers • Consumer feedback consideration
	10.	**Employee Concern** • Human rights • Working conditions • Health and Safety of employee • Employee’s training

Pre-processing involves conversion of .pdf files to the required .txt formats with Adobe Acrobat Pro software. After converting to .txt this files have been rechecked manually to ensure everything is in suitable format and ready for the next task. Each paragraph of these documents is considered as single unit input to the algorithm. All the paragraphs from each sustainability report is collected and divided into n small documents out of which any *n-1* is training set and the rest *1* is test document. Each paragraph is given a category out of the above mentioned 10 category either by manual observation in case of training set or computationally in case of test set. First word of each paragraph is be reserved for company name & year, followed by 2nd with category provided. Paragraphs with no category is given “00”. The stop words are removed from all the paragraphs using NLTK database. Redundant words will be replaced by words common for every company. For example, in each document all the names of industries is replaced by “company”, similarly with “year”. Generation of training is done by careful manual observation considering all the above mentioned rule.

Computerized Classification will be achieved using text mining and classification algorithm. **Text mining**, also referred as Text Analysis is the process of extracting useful information from unstructured textual data (Shahi et al., 2014). **Multinomial Naïve Bayes** will be used for classification**.** It is the modified version of basic Naïve Bayes algorithm. The reason behind using the algorithms is because it is very simple to implement and is very efficient. The Naïve Bayes classifier have been used in many text mining application due to its simple principle and high accuracy (Rish, n.d., 2001). Bayesian classifiers are based on the principle that presence or absence of a word in an article determines the outcome of the prediction (Lamkanfi et al., 2011). In these algorithm each processed term is given a probability that it belongs to a certain group. They are called naïve because of the assumption that no two words are related to each other unless they are same (Feldman & Sanger, 2006) which is obviously not true in the context of CSR reports. In this part we will automatically select paragraphs from test document and computationally assign them a category depending upon the previous experience by the algorithm on training data. This algorithm uses probabilistic measure to assign a category to the paragraphs based on the content of the paragraph.

The correctness measure we are using for the verification of the algorithms is called *F-measure* which is a combination of 2 indicators, 1) *Precision* & 2) *Recall.*

	Correct	**Not correct**
Selected	True positive **(tp)**	False positive **(fp)**
Not selected	False negative **(fn)**	True negative **(tn)**

Table 1: 2-by-2 contingency table

Precision tells us about the percentage of selected items that are correct, mathematically

$$Precision = \frac{true\ positive}{true\ positive + false\ positive}$$

And *Recall* tells us about percentage of correct items that are selected, mathematically

$$Recall = \frac{true\ positive}{true\ positive + false\ negative}$$

A combined measure that assesses the *Precision/Recall* tradeoff is F measure which is the weighted harmonic mean of both.

$$F\ measure = \frac{(\beta^2 + 1) * precision * recall}{(\beta^2 * precision) + recall}$$

Here β is the weighing measure between *precision* and *recall.*

If β=1, then *recall's* weight = *precision's* weight
Else β<1, then *recall's* weight < *precision's* weight

Manual Evaluation: A set of articles for which the category is already known, is needed to initialize the classifier. This is only possible by using manual evaluation. Assignment of category is done by careful observation of contents of each paragraph as per the above mentioned specification of each category. A minimum of 100 classified articles will be used in each category. This will be followed by semi-classified data for which categorization will be done in fast forward mode. The whole process is to generate more and more training set to increase the *F measure* of the system. It is expected that the accuracy will be very high, if around 500 training sets are used in each category.

Relation Extraction will be achieved by relaxing the Multinomial Naïve Bayes algorithm to an extent that instead of choosing the category with highest probability, now choose two categories with maximum and second maximum probability. It is intuitive that the *F measure* will increase in doing so but there is also possibility of increasing the error in classification. If we consider frequency of each pair of categories we can very easily plot a 2-Dimensional graph. In this relational graph strength of the relationships will be proportional to the frequency of appearing together in Multinomial Naïve Bayes classification.

RESULTS and ANALYSIS: Results have been categorized into three major categories 1) Performance of algorithm 2) Results on classification 3) Relationship extraction.

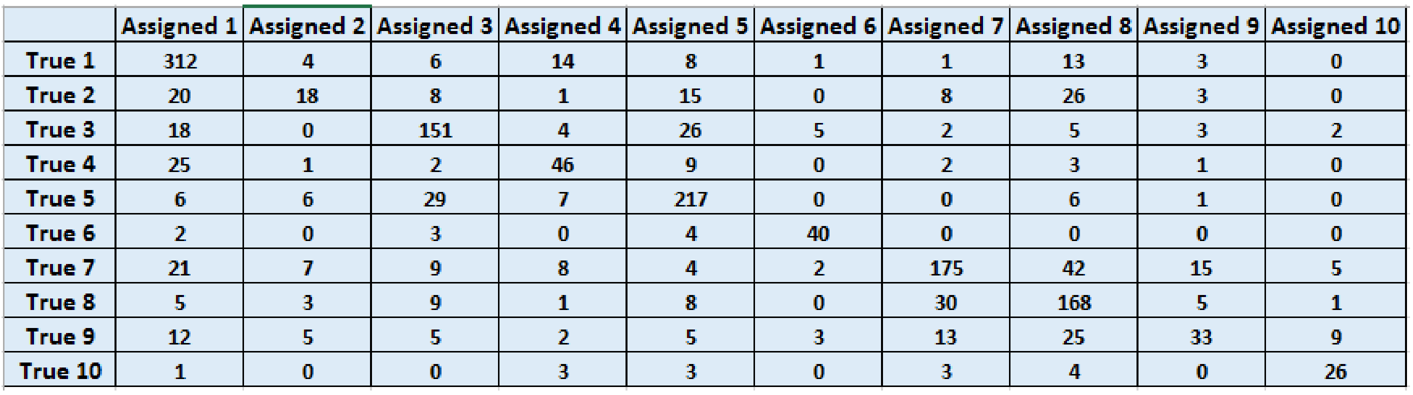
Performance of the Algorithm: The algorithm was trained on 5195 training article for which the classification is manually evaluated as discussed earlier. It was trained in a round robin fashion. Training was immediately followed by testing the algorithm on 1656 more articles. The training and test data was generated from Honda CSR reports (2000-2006). More generally speaking every time we will train our algorithm with 6 out of 7 CSR reports, and the remaining 7th is used as a test set. The output of the our algorithm is a 2D matrix called Confusion matrix $\boldsymbol{C}$ where number in $\boldsymbol{C_{ij}}$ represents the number of article which are actually of i^{th} category but our classifier classified it as an article of j^{th} category. Here number of correct classification are the diagonal elements of the Confusion matrix $\boldsymbol{C}$. The Confusion matrix created when evaluating the classifier with Honda CSRs is given in Table 1. The Confusion matrix is also used in calculation of recall and precision value for all the categories individually.

Recall value for a category i is given by $R(i)$ where

$$R(i) = \frac{C_{ii}}{\sum_j C_{ij}}$$

Similarly *Precision* value for a category is given by $P(i)$ where

$$P(i) = \frac{C_{ii}}{\sum_j C_{ji}}$$



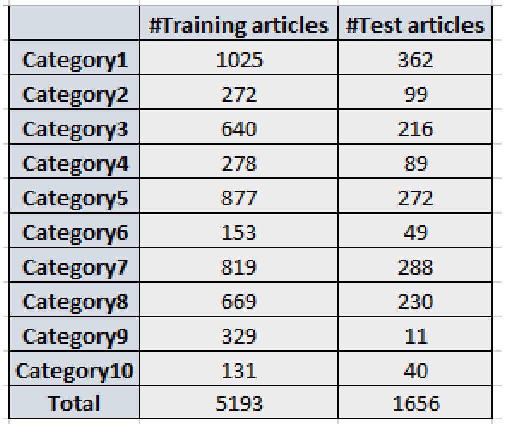
	Assigned 1	Assigned 2	Assigned 3	Assigned 4	Assigned 5	Assigned 6	Assigned 7	Assigned 8	Assigned 9	Assigned 10
True 1	312	4	6	14	8	1	1	13	3	0
True 2	20	18	8	1	15	0	8	26	3	0
True 3	18	0	151	4	26	5	2	5	3	2
True 4	25	1	2	46	9	0	2	3	1	0
True 5	6	6	29	7	217	0	0	6	1	0
True 6	2	0	3	0	4	40	0	0	0	0
True 7	21	7	9	8	4	2	175	42	15	5
True 8	5	3	9	1	8	0	30	168	5	1
True 9	12	5	5	2	5	3	13	25	33	9
True 10	1	0	0	3	3	0	3	4	0	26

Table 1: Confusion matrix for the classifier with Honda CSRs (2000-06)

The calculated recall and precision value for all the categories is given in Table 3. The average in the Table 3 is obtained by macro-averaging the values from each category. Macro-averaging was chosen over micro-averaging to give equal importance to all the categories. The average recall obtained is 61.56 and average precision value is 63.18. Another measure called accuracy *Acc* which is given by sum of correct classification divided by total number of articles tested for classification. Mathematically

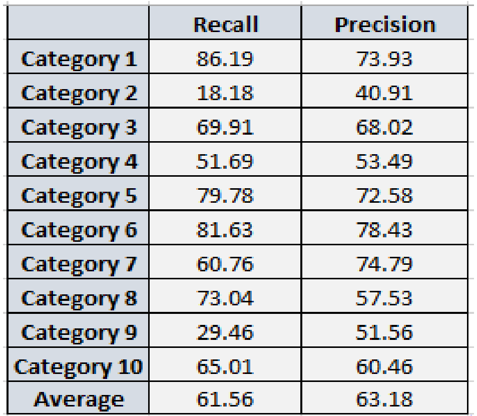
$$Acc = \frac{\sum_i C_{ii}}{\sum_i \sum_j C_{ij}}$$

The accuracy measure for this set was found to be 67.39. The number of articles for both training and test set for all the categories can be found on Table 2. Our classifier classified 1115 out of 1656 test articles correctly. The reason behind low recall value of category 2 is because the articles containing supply chain management topics get misclassified into 4 R's, emission and green building. This is because supply chain management article contains information regarding reduction of packaging materials, better logistics which reduce carbon emission. Also as the new green buildings are coming up, article containing this kind of information can easily fall into Business Development (Category2) or Green manufacturing (Category 8). Also as the size of the training corpus increase over time we can expect better performance of the algorithm.



	#Training articles	#Test articles
Category1	1025	362
Category2	272	99
Category3	640	216
Category4	278	89
Category5	877	272
Category6	153	49
Category7	819	288
Category8	669	230
Category9	329	11
Category10	131	40
Total	5193	1656

Table 2: Count of number of articles in training and test sets.



	Recall	Precision
Category 1	86.19	73.93
Category 2	18.18	40.91
Category 3	69.91	68.02
Category 4	51.69	53.49
Category 5	79.78	72.58
Category 6	81.63	78.43
Category 7	60.76	74.79
Category 8	73.04	57.53
Category 9	29.46	51.56
Category 10	65.01	60.46
Average	61.56	63.18

Table 3: Recall and precision value for each category

Result on Classification: When we plot the classification result of each document after normalizing with the total number of articles we have in that particular document we see the graph shown in Fig 1. In Fig 1 each point in *x-axis* represents 1 of the categories of classification and similarly the value shown in *y-axis* are the percentage measure of each category in the corresponding document. The basic assumption we are considering here is that if an industry is talking more about a particular sustainability concern that means it is spending much of its resources and effort in that particular area. Also not every have the same reputation in the market and concerns towards sustainability add a lot of value towards company's reputation which ultimately lead to increase in sales, number of loyal customers and also external interest in company's stock.

Here Fig 1 is obtained after averaging the value obtained from Honda CSR reports in the year range 2000-2006. Fig 1 clearly represents the involvement of a Honda in each of the sustainability area explicitly. Looking closely at Fig 1 we could see that very less proportion of Honda CSR reports focuses in areas like *Noise Control, Employee concern, Waste Disposal, Business Development and Supply Chain Management* but it mentioned a lot about *4 R's (Reuse, Reduce, Recycle, Remanufacture), Emission control, Community Involvement and Award.* The imbalance in distribution of concern and budget is definitely a point of concern for the stakeholders and graph like this can be very helpful for governments to regulate, researchers to study behavior of an industry.

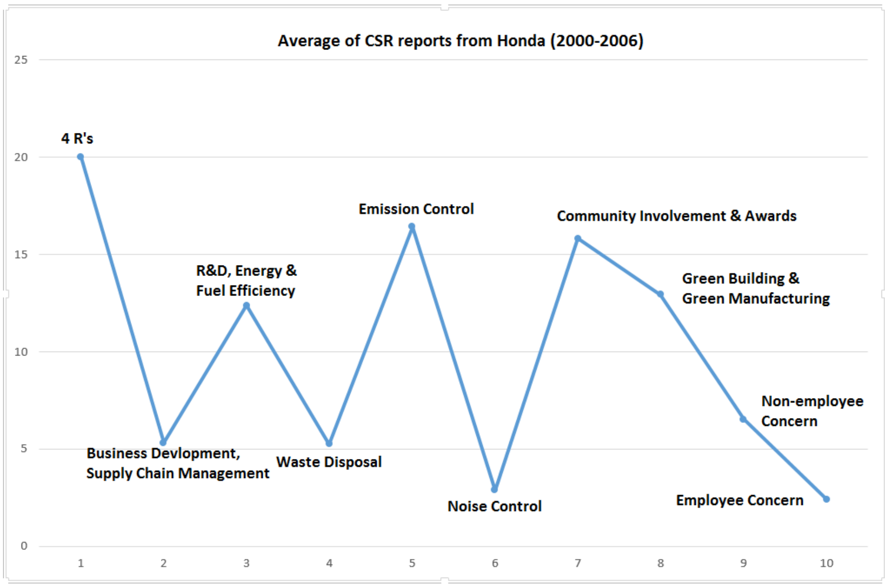


Fig 1: Average analysis of Honda CSR reports (2000-06)

Though the above graph gave an insight into Honda's concern towards corporate social responsibility but it says nothing about the dynamic change in nature. So we plotted the graph obtained for different year together at single location and obtained Fig 2. Fig 2 depicts the variation of sustainability concern of Honda over the years 2000-2006 and we can see that there is not much change happening in the considered 7 years, expect for the Emission part. The content of CSR in 2002-2003 of Honda contains a lot of detail regarding reduction in emission. The concern regarding waste disposal and Noise reduction are very low in all of the years and no major fluctuation has occurred in the years. This similarity made us to look closely at these CSR reports and what we found was a lot of repetition of data from the previous years. By repetition we mean that the content of CSR report of previous years are directly copied in CSR reports of later years. What they usually do is if they do an environmental project in year 2000, they mentioned it in every following CSR reports. This implies that not much new work is happening in the area of sustainability of the industry and they are caring forward the old projects and repeatedly mentioning them in new CSR reports to make it large and content full.

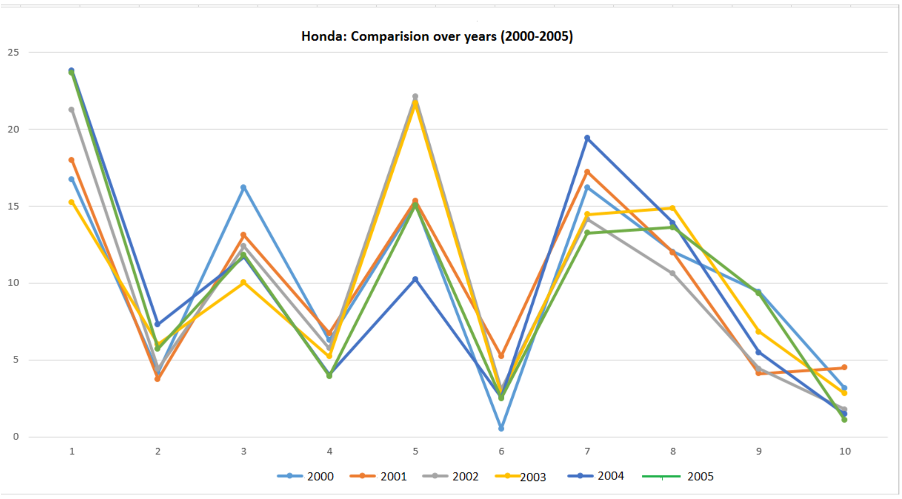


Fig 2: Comparison of CSR reports from Honda (2000-2005)

We also compared the results of various industries with each other and one such graph is shown in Fig 3. While drawing Fig 3 we have considered CSR reports from Bosch (2003-04), Toyota (2005-06), Volkswagen (2009) and Fiat (2007-08). The first result we obtained from this observation that CSR or Sustainability means different to different industry. In Fiat CSR (2007-08) approximately 30% content is on expansion of business and improving supply chain management where as in contrary Toyota CSR (2005-06) only 10% content is on business development and supply chain improvement. Graph like this could be of great importance while taking various decisions. For example, if you are an employee and want the safer and peaceful work environment then among the 4 mentioned in Fig 3, Volkswagen will be the best option. Similarly if we want a car which is better at fuel efficiency then we Toyota will be the best choice. For a governmental institution which regulates the waste disposal of these industries, they can very easily see that Fiat and Volkswagen is showing very less concern towards it and they should be the first one to put a check on. Similarly we can extract various kind of useful information from Fig 3 depending upon the situation and the requirement. The above mentioned results can vary from actual results because these results are completely based on content of CSR reports. If every industry provide complete and correct information in their CSR reports these graph can be of very use and will give us precise data without much effort.

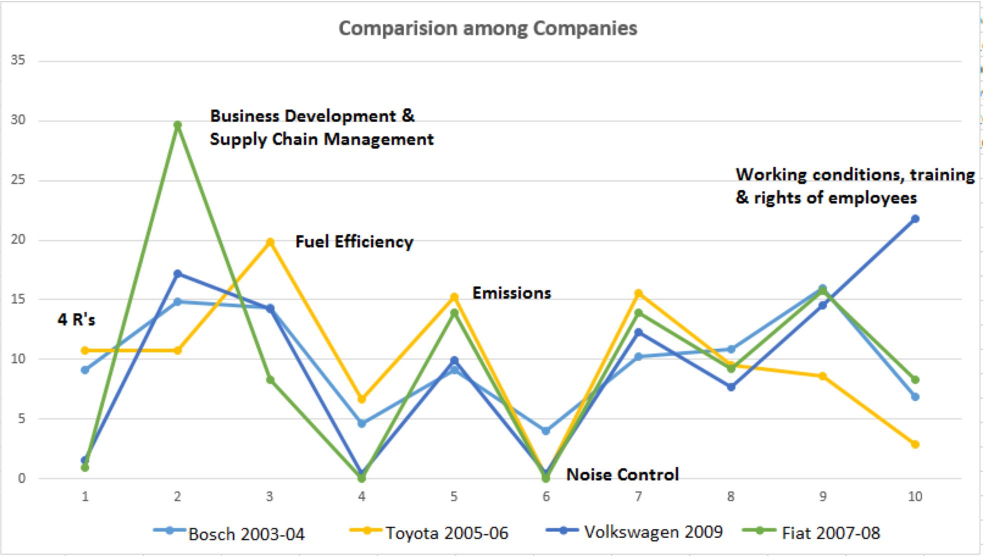


Fig 3: Comparison of CSR reports from 4 different industries

Result on Relationship extraction: When we modified our algorithm to choose 2 categories for each article instead of 1 we got an accuracy $Acc2$ of 81.92 which is a significant improvement but involves a lot of error. But the basic motivation behind doing this was not to improve accuracy value but to extract relationship between the above mentioned sustainability areas. To draw the relationship curve we calculated the frequency of co-occurrence of various area together after classification. In the graph the visual measure we have used is the thickness of the edge between 2 nodes, this thickness is directly proportional to the strength of the relationship or in other terms frequency of co-occurrence. In Fig 4 you can see that the number just above any edge gives us the co-occurrence of the end nodes connecting this edge. Also the number at each nodes denotes the category number of each sustainability area. The relationship graph gave us the interdependence of various sustainability areas. For example, Fuel, Energy Efficiency and emission co-occurred 287 times in the classification showing strong interdependence between them. The explanation of this relationship is very simple more fuel and energy efficient less fuel combustions which will imply to less emission. Similarly thick edge between waste disposal (Node 4) and 4 R's (Node 1) is because more an industry is focusing on reusing, reducing the use of material less waste is what it is producing. This graph can be useful to those industries who want to distribute their limited resources on the sustainability areas. Moreover we can see that if we work on fuel efficiency we will automatically get better results in emissions reduction.

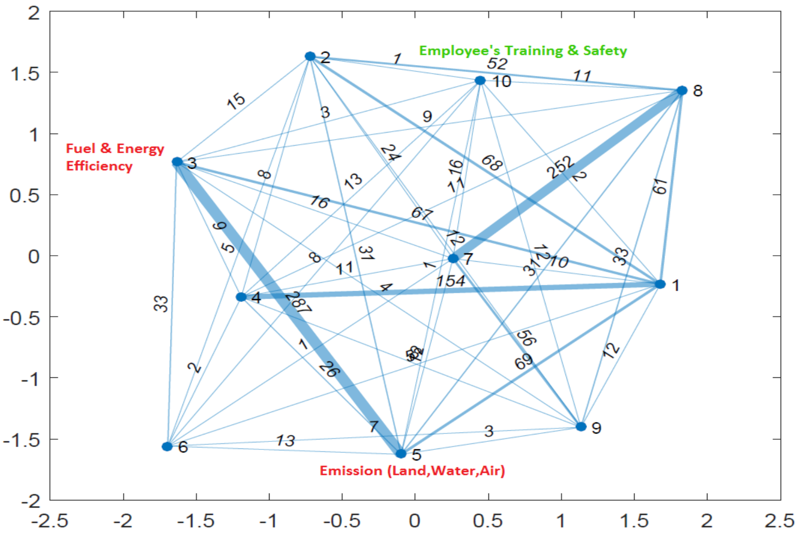


Fig 4: Relationship between various sustainability Areas obtained after altering Multinomial Naïve Bayes

DISCUSSION and CONCLUSION: This study provides an easy and novel way to look into large CSR reports from automotive industry. This study is based on the hypothesis that if a company is writing more about a particular area of sustainability then it is more concerned about that particular area. This study provides a novel way to represent company's concern about sustainability in a pictorial way. This pictorial way also provides a method to look at sustainability trends over the years. In this study we have taken 7 reports from Honda to analyse this, but this can be very easily extended to as many papers as we want. Using this techniques we can very easily visualize the sustainability trends over longer period of time to draw general results independent of company. More inter-industry comparison helps to find what other competitor is working on. This research will also help entrepreneur or new industries to allocate their limited resources in the best possible way. One result that we have seen from the Honda intra-industry comparison curve is that repeated mentioning of data is happening in CSR reports which is increasing the size of the reports and also providing misleading information to the consumer. Governments, NGOs or other interested authority must look into the matter to verify their sustainability concern. Moreover the direct implication of this research to the real world will be fruitful if company provide true and complete information in CSR reports.

The inter-relationship graph Fig 4 shows that various areas of sustainability can be merged into a single area as one is a direct implication of other. This research also provides insight that what factors are really important in the existence and growth of a company. This study showed that currently there exists no standard pattern of amount of concern a company should put in sustainability. It showed that concern towards sustainability depends on the market situation and the demand of stakeholder.

LIMITATION AND FUTURE SCOPES: The limitations that we have in current research is that it involves supervised learning algorithm which involves manual effort to make the training set. More over Multinomial Naïve Bayes algorithm assumes that all terms occur independent of each other which is certainly not true (Feldman & Sanger, 2006). Also the in the data we have used the distribution of number of articles is not uniform and some categories do not have sufficient number of articles. This scarcity of data is giving the poor performance measure of the algorithm. A future scope is to use and compare various algorithms like classification algorithm like Support Vector Machine, K mean etc. Moreover it will be beneficial to take this forward to complete automatic process which involves no manual evaluation. Application of centering resonance analysis (CRA) to each category to find inter and intra category relations will also be an interesting research area. Another future scopes will be to build a computerized tool which can automatically extract information and analyse it.

REFERENCES:

Amrina, E., & Yusof, S. M. M. (2011). Key performance indicators for sustainable manufacturing evaluation in automotive companies. *Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on*, 1093–1097. http://doi.org/10.1109/IEEM.2011.6118084

Chae, B. K. (2015). Int . J . Production Economics Insights from hashtag # supplychain and Twitter Analytics : Considering Twitter and Twitter data for supply chain practice and research. *Intern. Journal of Production Economics*, *165*, 247–259. http://doi.org/10.1016/j.ijpe.2014.12.037

Commission, W. (n.d.). Report of the World Commission on Environment and Development : Our Common Future Acronyms and Note on Terminology Chairman ' s Foreword.

Compact, U. N. G. (2011). Towards a New Era of Sustainability in the Automotive Industry Contents.

Du, S., Bhattacharya, C. B., & Sen, S. (2010). Maximizing business returns to corporate social responsibility (CSR): The role of CSR communication. *International Journal of Management Reviews*, *12*(1), 8–19. http://doi.org/10.1111/j.1468-2370.2009.00276.x

Feldman, R., & Sanger, J. (2006). *The text mining handbook* (Vol. 1). http://doi.org/10.1017/CBO9781107415324.004

GRI Sustainability Reporting Statistics Publication year 2011. (2011), (April).

Jose, A., & Lee, S. M. (2007). Environmental reporting of global corporations: A content analysis based on Website disclosures. *Journal of Business Ethics*, *72*(4), 307–321. http://doi.org/10.1007/s10551-006-9172-8

Kolk, A. (2003). Trends in sustainability reporting by the Fortune Global 250. *Business Strategy and the Environment*, *291*(12), 279–291. http://doi.org/10.1002/bse.370

Kurucz, E. C., Colbert, B. A., & Wheeler, D. (2008). The Business Case for Corporate Social Responsibility. *The Oxford Handbook on Corporate Socail Responsibility*, 83–112.

Lamkanfi, A., Demeyer, S., Soetens, Q. D., & Verdonckz, T. (2011). Comparing mining algorithms for predicting the severity of a reported bug. *Proceedings of the European Conference on Software Maintenance and Reengineering, CSMR*, 249–258. http://doi.org/10.1109/CSMR.2011.31

Mahoney, L. S., Thorne, L., Cecil, L., & LaGore, W. (2013). A research note on standalone corporate social responsibility reports: Signaling or greenwashing? *Critical Perspectives on Accounting*, *24*(4-5), 350–359. http://doi.org/10.1016/j.cpa.2012.09.008

McDonald, D., McNicoll, I., Weir, G., Reimer, T., Jacobs, N., & Bruce, R. (2012). The Value and Benefit of Text Mining. *JISC Digital*, (March), 1–32. Retrieved from http://bit.ly/jisc-textm

Modapothala, J. R., & Issac, B. (2009). Study of Economic , Environmental and Social Factors in Sustainability Reports using Text Mining and Bayesian Analysis, (Isiea), 209–214.

Nishant, R., Goh, M., & Kitchen, P. J. (2015). Sustainability and differentiation : Understanding materiality from the context of Indian fi rms ☆ , ☆☆. *Journal of Business Research*, *69*(5), 1892–1897. http://doi.org/10.1016/j.jbusres.2015.10.075

Panayiotou, N. A., & Aravossis, K. G. (2009). Greece: A comparative study of CSR reports. *Global Practices of Corporate Social Responsibility*, (December), 1–508. http://doi.org/10.1007/978-3-540-68815-0

Rish, I. (n.d.). An empirical study of the naive Bayes classifier, 41–46.

Rivera, S. J., Minsker, B. S., Work, D. B., & Roth, D. (2014). Environmental Modelling & Software A text mining framework for advancing sustainability indicators. *Environmental Modelling and Software*, *62*, 128–138. http://doi.org/10.1016/j.envsoft.2014.08.016

Rolland, D., Keefe, J. O., & Bazzoni, J. O. K. (2010). Greening corporate identity : CSR online corporate identity reporting. http://doi.org/10.1108/13563280910980041

Saha, A., & Nabareseh, S. (2015). Communicating Corporate Social Responsibilities : Using Text Mining for a Comparative Analysis of Banks in India and Ghana, *6*(3), 11–20. http://doi.org/10.5901/mjss.2015.v6n3s1p11

Seuring, S., & Muller, M. (2008). From a literature review to a conceptual framework for sustainable supply chain management. *Journal of Cleaner Production*, *16*(15), 1699–1710. http://doi.org/10.1016/j.jclepro.2008.04.020

SHAHI, A. M., ISSAC, B., & MODAPOTHALA, J. R. (2014). Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring. *International Journal of Computational Intelligence and Applications*, *13*(01), 1450006. http://doi.org/10.1142/S1469026814500060

Soiraya, B. (2011). Semi-automatic Green ICT Ontology Construction from CSR Report, 711–714.

Sukitsch, M., Engert, S., & Baumgartner, R. J. (2015). The implementation of corporate sustainability in the European automotive industry: An analysis of sustainability reports. *Sustainability (Switzerland)*, *7*(9), 11504–11531. http://doi.org/10.3390/su70911504

Tate, W. L., Ellram, L. M., & Kirchoff, J. F. (2010). Corporate social responsibility reports: A thematic analysis related to supply chain management. *Journal of Supply Chain Management*, *46*(1), 19–44. http://doi.org/10.1111/j.1745-493X.2009.03184.x

Te, W., Adhitya, A., & Srinivasan, R. (2014). Computers in Industry Sustainability trends in the process industries : A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400. http://doi.org/10.1016/j.compind.2014.01.004

Tremblay, M. C., Parra, C. M., & Castellanos, A. (2016). Corporate Social Responsibility Reports : Understanding Topics via Text Mining, (August 2015).

Vlachos, P. A., & Tsamakos, A. (2009). Corporate social responsibility : attributions , loyalty , and the mediating role of trust, 170–180. http://doi.org/10.1007/s11747-008-0117-x

Wati, Y., & Koo, C. (2010). The Green IT Practices of Nokia , Samsung , Sony , and Sony Ericsson : Content Analysis Approach, 1–10.

Worcester, R. (2007). Reflections on corporate reputations. http://doi.org/10.1108/00251740910959422

Amrina, E., & Yusof, S. M. M. (2011). Key performance indicators for sustainable manufacturing evaluation in automotive companies. *Industrial Engineering and Engineering Management (IEEM), 2011 IEEE International Conference on*, 1093–1097. http://doi.org/10.1109/IEEM.2011.6118084

Chae, B. K. (2015). Int . J . Production Economics Insights from hashtag # supplychain and Twitter Analytics : Considering Twitter and Twitter data for supply chain practice and research. *Intern. Journal of Production Economics*, *165*, 247–259. http://doi.org/10.1016/j.ijpe.2014.12.037

Commission, W. (n.d.). Report of the World Commission on Environment and Development : Our Common Future Acronyms and Note on Terminology Chairman ' s Foreword.

Compact, U. N. G. (2011). Towards a New Era of Sustainability in the Automotive Industry Contents.

Du, S., Bhattacharya, C. B., & Sen, S. (2010). Maximizing business returns to corporate social responsibility (CSR): The role of CSR communication. *International Journal of Management Reviews*, *12*(1), 8–19. http://doi.org/10.1111/j.1468-2370.2009.00276.x

Feldman, R., & Sanger, J. (2006). *The text mining handbook* (Vol. 1). http://doi.org/10.1017/CBO9781107415324.004

GRI Sustainability Reporting Statistics Publication year 2011. (2011), (April).

Jose, A., & Lee, S. M. (2007). Environmental reporting of global corporations: A content analysis based on Website disclosures. *Journal of Business Ethics*, *72*(4), 307–321. http://doi.org/10.1007/s10551-006-9172-8

Kolk, A. (2003). Trends in sustainability reporting by the Fortune Global 250. *Business Strategy and the Environment*, *291*(12), 279–291. http://doi.org/10.1002/bse.370

Kurucz, E. C., Colbert, B. A., & Wheeler, D. (2008). The Business Case for Corporate Social Responsibility. *The Oxford Handbook on Corporate Socail Responsibility*, 83–112.

Lamkanfi, A., Demeyer, S., Soetens, Q. D., & Verdonckz, T. (2011). Comparing mining algorithms for predicting the severity of a reported bug. *Proceedings of the European Conference on Software Maintenance and Reengineering, CSMR*, 249–258. http://doi.org/10.1109/CSMR.2011.31

Mahoney, L. S., Thorne, L., Cecil, L., & LaGore, W. (2013). A research note on standalone corporate social responsibility reports: Signaling or greenwashing? *Critical Perspectives on Accounting*, *24*(4-5), 350–359. http://doi.org/10.1016/j.cpa.2012.09.008

McDonald, D., McNicoll, I., Weir, G., Reimer, T., Jacobs, N., & Bruce, R. (2012). The Value and Benefit of Text Mining. *JISC Digital*, (March), 1–32. Retrieved from http://bit.ly/jisc-textm

Modapothala, J. R., & Issac, B. (2009). Study of Economic , Environmental and Social Factors in Sustainability Reports using Text Mining and Bayesian Analysis, (Isiea), 209–214.

Nishant, R., Goh, M., & Kitchen, P. J. (2015). Sustainability and differentiation : Understanding materiality from the context of Indian firms ☆ , ☆☆. *Journal of Business Research*, *69*(5), 1892–1897. http://doi.org/10.1016/j.jbusres.2015.10.075

Panayiotou, N. A., & Aravossis, K. G. (2009). Greece: A comparative study of CSR reports. *Global Practices of Corporate Social Responsibility*, (December), 1–508. http://doi.org/10.1007/978-3-540-68815-0

Rish, I. (n.d.). An empirical study of the naive Bayes classifier, 41–46.

Rivera, S. J., Minsker, B. S., Work, D. B., & Roth, D. (2014). Environmental Modelling & Software A text mining framework for advancing sustainability indicators. *Environmental Modelling and Software*, *62*, 128–138. http://doi.org/10.1016/j.envsoft.2014.08.016

Rolland, D., Keefe, J. O., & Bazzoni, J. O. K. (2010). Greening corporate identity : CSR online corporate identity reporting. http://doi.org/10.1108/13563280910980041

Saha, A., & Nabareseh, S. (2015). Communicating Corporate Social Responsibilities : Using Text Mining for a Comparative Analysis of Banks in India and Ghana, *6*(3), 11–20. http://doi.org/10.5901/mjss.2015.v6n3s1p11

Seuring, S., & Muller, M. (2008). From a literature review to a conceptual framework for sustainable supply chain management. *Journal of Cleaner Production*, *16*(15), 1699–1710. http://doi.org/10.1016/j.jclepro.2008.04.020

SHAHI, A. M., ISSAC, B., & MODAPOTHALA, J. R. (2014). Automatic Analysis of Corporate Sustainability Reports and Intelligent Scoring. *International Journal of Computational Intelligence and Applications*, *13*(01), 1450006. http://doi.org/10.1142/S1469026814500060

Soiraya, B. (2011). Semi-automatic Green ICT Ontology Construction from CSR Report, 711–714.

Sukitsch, M., Engert, S., & Baumgartner, R. J. (2015). The implementation of corporate sustainability in the European automotive industry: An analysis of sustainability reports. *Sustainability (Switzerland)*, *7*(9), 11504–11531. http://doi.org/10.3390/su70911504

Tate, W. L., Ellram, L. M., & Kirchoff, J. F. (2010). Corporate social responsibility reports: A thematic analysis related to supply chain management. *Journal of Supply Chain Management*, *46*(1), 19–44. http://doi.org/10.1111/j.1745-493X.2009.03184.x

Te, W., Adhitya, A., & Srinivasan, R. (2014). Computers in Industry Sustainability trends in the process industries : A text mining-based analysis. *Computers in Industry*, *65*(3), 393–400. http://doi.org/10.1016/j.compind.2014.01.004

Tremblay, M. C., Parra, C. M., & Castellanos, A. (2016). Corporate Social Responsibility Reports : Understanding Topics via Text Mining, (August 2015).

Vlachos, P. A., & Tsamakos, A. (2009). Corporate social responsibility : attributions , loyalty , and the mediating role of trust, 170–180. http://doi.org/10.1007/s11747-008-0117-x

Wati, Y., & Koo, C. (2010). The Green IT Practices of Nokia , Samsung , Sony , and Sony Ericsson : Content Analysis Approach, 1–10.

Worcester, R. (2007). Reflections on corporate reputations. http://doi.org/10.1108/00251740910959422