Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 03, March-2023

Sentimental ESG Mining using Machine Learningand AI Techniques

Pritish Sinha
Computer Science and Engineering
Galgotias University Greater Noida, India

Khushi
Computer Science and Engineering
Galgotias University Greater Noida, India

Siddhant Jain
Computer Science and Engineering
Galgotias University Greater Noida, India

Vijaya Chauodhary
Computer Science and Engineering
Galgotias University Greater Noida, India

***Abstract*—Environment, Social & Governance are new generalised terms in sustainability of businesses and organisations. ESG Policies are now imposed in good financing terms and business development. Many companies has followed sustainable paths while others are now turning their way through this. New investors seek for information regarding ESG of an investment so to keep their capital safe and keep developing with new regulations of law makers. This calls for large committee and decision making with including proper documentations of each aspect in company working from electricity consumption to policies that safeguards human rights, thus each organisation maintains documents namely CSR Report, Annual Report, Social-Governance Documents etc however an investor doesn't directly reads these documents to factor their investment. They hire business analyst companies and their ESG analyst goes through manual task of reading each entry and filing a compiled factset with reviews according to global standards. Each organisation have different ways of documenting their data. Thus, text sentimental mining comes to play role here and reduce task of compiling and reviewing these large data scaling down each process time to less than hours of work which took more than days in traditional method.**

***Index Terms*—ESG, factset, visualisation, sentiment analysis**

I. Introduction

There are number of companies that have setup it's governance to measure, and rate it's sustainability. S&P 500 companies have board level committees to look over this aspect of business and many hire analyst companies that keep track on their annual reports and documents related to social and governance. In this paper we are examining disclosure of these data to machine learning model and AI model that will be analysing sentimental values of data to rate a business model over environment,social and governance topics.

A. *Problem Statement*

ESG analyst work is of more than days depending on how data is documented by an organisation, scale of investment in organisation, number of factory sites handled by organisation, other investment of that organisation etc. Additionally, green washing is a problem which is falsely using of data and words to trick an analyst to come to conclusion that organisation is working under right full terms. Another disadvantage of this manual labour is that data is not timestamped and not all real time data being evaluated.

B. *Problem Solution*

After some research it can be concluded that many data segments of these organisation are being published regularly at sites like GDelt, Yahoo and other. We can make use of machine learning and AI techniques that can help to visualise data of organisation and also provide factored filter option. Proposed paper plans to use machine learning, concept of NLP to extract sentimental features of data and scale them to visualise whole work. Secondary objective of this work is to provide data of connected/related organisation and better investment options to it's user using Node2vec embedded connection graphs.

II. Literature Survey

A. *Background Research*

There's been increase in governance regulation to many organisation to follow ESG standards in heed to sustainable development. Smart investment is now subject to investment in firms/organisation that is following global standards and UNSDP regulation in 3 terms 'E', 'S' & 'G' i.e, Environment, Social and Governance. We have examined disclosure of ESG activity data by organisation to prove there ESG standards and actract the investors. [10]

However there are many cases of malpractices in area of research and analysis as it is tricky and only based on data released by firms. One of these is greenwashing that is publishing data that makes company looks like it is following and aiming to work upon SD path.

In other words such laws are costly as pooling the whole disclosure can lead to destruction of share holder value. We have also found number of investment in analyst groups and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 03, March-2023

firms that work on contract basis to pool and analyse data by manually working on data like CSR report, Annual report and Social-Governance sheets and policies chart.

B. *Related Work*

Sentiment analysis and opinion mining field has grown over last few years. Are of research now aims large set of application in real life for example use of reviews from customer to enhance a product, enhance users accessibility of product use like spotify, instagram, facebook, web browsing, etc. These all applications are utilizing higher artificial intelligence techniques to determine users feelings, opinion and natural language processing is at boss to solve these urges. Many scientific articles are being published and worked upon related to opinion/ emotion mining from text data like reviews, comments and tweets. [12]

This section of paper aims to provide a curated list of existing models, implementation and other related work. Many dictionary types like WordNet developed SentiWordNet [13] these generalise terms related to emotions but major drawback is that these dictionary types cannot implement hierarchical word net failing proper testiment of sentiment analysis. Below is table of obtained works with acronyms 'ML' to represent machine learning work, 'L' for lexicon and 'H' to denote hybrid work.

TABLE I
RELATED WORKS

Year	Approach	Domain	Result
2002	ML	Movie Review [14]	82.9%
2009	ML	Product Review [15]	83.30%
2013	L	Word Dictionary [7]	93.30%
2015	ML	Tweet [5]	67.40%
2017	L	Review [10]	83.30%
2018	H	Customer Review [8]	83.30%
2019	H	General [4]	83.30%
2020	H	ESG - Japenese [13]	83.30%
	ML	ESG- Spanish,English [12]	91.2%

III. VISUALIZATION MODEL

Model summary includes steps involved in development which entails: preprocessing, word embedding, classification, and visualization. Below are explanation of each step.

A. *Step 1- Preprocessing*

Following the nature of making a machine learning model we collected our data first for preprocessing. We gathered over 120+ CSR, Annual, Social-Governance and Policy Documents reports that followed US and French standards, formatted each pdf and extracted proper nouns, action verbs and selective 'E', 'S'& 'G' dictionary texts ex- sewage, coal, monoxide, accident, crash, etc while manually stemmed plurals in more than few cases.

TABLE II
CLASSIFIED WORD IN 'E', 'S', & 'G" CATEGORY

Category	Word Details	
it falls	*Word*	*Count*
E - Environment	global warming, wastage	1573
S - Social	overtime, violation	798
G - Governance	committee, succession plan	1360
Other	profit, data, activity	24672
Total		28403

B. *Step 2- Word Embedding*

In proposed paper we have used word2vec neural mapping model to generate out word embedding vector. This has generated a 50-dimensional vector of word embedding, for specificity skip gram model of word2vec [7] model has been employed to structure a hierarchical tree for word embedding.

C. *Step 3- Word Classification*

Our classification basket consist of 4 categories which are 'Environment', 'Social', 'Governance', & 'Other' embedded in word vector label of 'E', 'S', 'G' & 'O' respectively. Some of labeling done manually through local natural language toolkit directed dictionary for ESG word stack and rest repetitive words labeled as Other, remaining classification is handled by neural network. Outcome of this is 1573 E-categorised word, 798 S- categorised words and 1360 G- categorised words rest of words are labeled as O here. Use of mono words is not beneficial so word structuring aided interpretation of statements this can be done by adding hierarchical words on basis of frequency and word in tree model. Example "Water Pollution" where 'Pollution' is 'Environment' topic and "Waste Diffuse", "Insoluble impurities" are required to be included in 'Environment' topic also. Word structuring

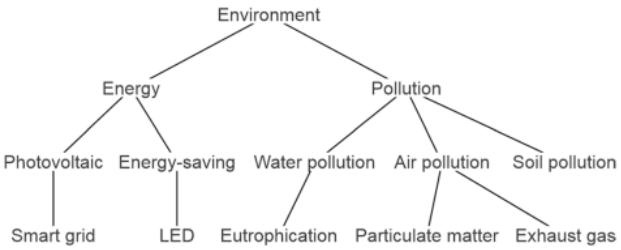

Fig. 1. Example of word structuring

requires indexing proper nouns of highlighted topics of ESG i.e, for "environment","social", and "governance" (k=1) fixed and for diverse words we make a word embedding vector v_i and v_j. We have d_{ij} as

$d_{ij} = (1 - cos(v_i, v_j))^2$

The value of cos ranges from -1 to 1, and according to calculation d_{ij} falls under range 0 to 4. But tree structure sum

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 03, March-2023

of divergence cannot be constructed as heuristic foundation is already established. Tree structure optimization is the key which can be worked on in future for better word structuring and we can apply other constraint algorithms for solution like greedy algorithm.

TABLE III
WORD COUNT

Word	*Count*
Absolute zero target	73
Cap and trade	25
CO_2	14
carbon emissions	13
DEI	0
EHS	0
EPA	4
Glasgow COP2	6
GHG	4
Scope 1	3
Scope 2	5
Scope 3	7
LCA	2
Net-Zero carbon emission	24
Offset	2
Paris Agreement	2

Our approach is non-optimised based for large set of words. Each tree can cover hundreds of words but presented solution will create similar sets of trees with quantitative word frequency. For example take instance from below image as process start discard a branch taking reference from Fig. 1 we used above.

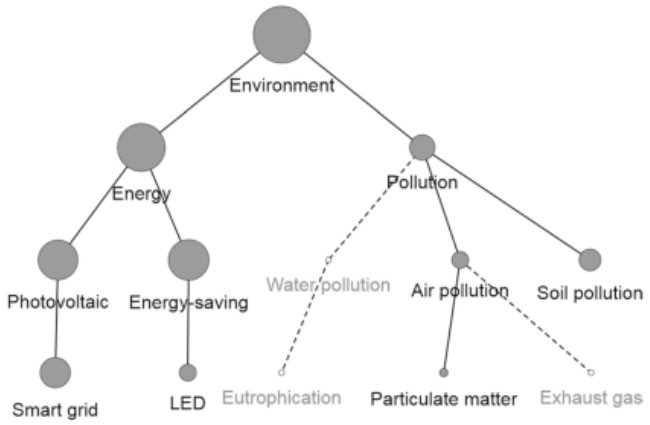

Fig. 2. Example of word frequency

D. *Step 4- Visualisation*

From Table 2 and Figure 2 we form a hypothesis related to E categorised words. In figure 2 darker the dots are and the diameter represents word frequency. Thus we can discern that organisation is somewhat related to Energy consumption and either there is bad impact on environment or it's detailing good environment friendly techniques leading to lower air and soil pollution. White dots are one that are discarded due to too low or zero frequency utilised from nlk tool dictionary.

*1) **Visualisation Summary**:* Figure 3 is depiction of possible tree result from an Annual Report following S&P global standards in ESG sustainable development and quality drive in it's good social and governance regulations. Connection of category E is formed here and shows that firms engages to issues of ecosystem with rare, indigenous species involved in it, on right hand side we see there are some topics related to global warming, discharge in air in relation to energy consumption or generation.

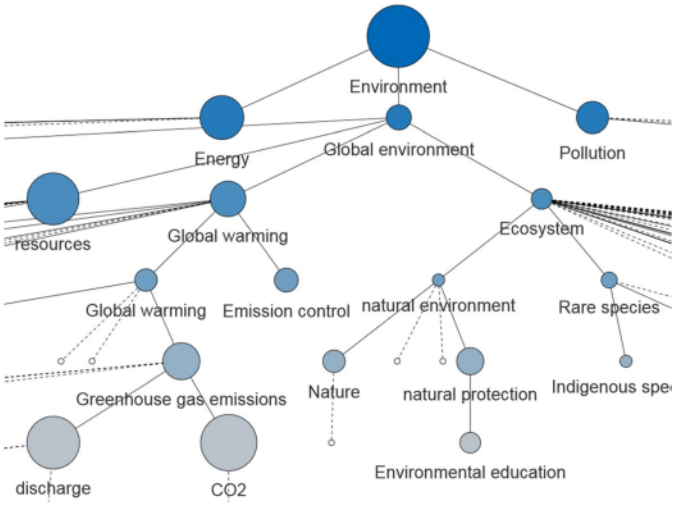

Fig. 3. Example of final word frequency

IV. ESG SCORING

Main objective of proposed paper is to provide feature of rating an organisation on basis of ESG activities and provide a comparison foundation of it. Proposed solution is to do a sentimental analysis on extracted data over selective topics and then set a parameter to score these analysis, use of emotion artificial intelligence, natural language processing(NLP) technique is used here.

*1) **What is Sentiment Analysis**:* Sentiment analysis falls under natural learning processing(NLP) technique that is either subjective analysis or opinion mining on basis of identified patterns of large data. It generates score card according to attitude, views of the feeded topic. [3]This analysis is structured over 3 elements of context:

- **Opinion/Emotion**: Opinion refers to polarity of subject while emotion scales qualitative features of sad, joy, anger and happiness.
- **Subject**: In sentiment analysis it's a crucial leg to provide subject over which analysis is targeted to get better and accurate ratings.
- **Organisation/Holder**: In sentiment analysis we should be providing scale of passiveness to feature evaluation for example reference of author who expressed and to whom it is expressed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 03, March-2023

A. Scoring function

To score a firm according to it's ESG activities we can't be using single function to define as a whole. [9]For this we define separate specificity proxies for each category and then the solid quality score $s^{q}nt_{i,t}$ it is the sum of logarithms of $n_{k,i,t}$ for each frequency count heads in generalised topics of word *K* in time *t* and of firm reports *i*. Sum is by default 1 so we have added 1 to $n_{k,i,t}$ by general, preventing it from being *0*.

$$s_{i,t}^{qnt} = \sum_{k=1}^{k} \log_n(n_{k,i,t}+1)$$

Industrial average $s_{i,t}^{spc}$ is average of total divergence to number of establishment transition in words of 'E', 'S' and 'G'. This average function is calculated because there can be number of different establishment or investment of a firm, these other variables will have different ESG activities and can factor the scoring for example- Firm 'A' have total 3 plants*('I', 'O' & 'U')* in different sectors of market now A uses plant I for electricity generation, O is sewage plant and U is manufacturing unit of A. E related activities is possible in all three plants, S and G is limited to plant U as this plant is manual operation i.e, we have to score each plant to consider ESG activity of A. For this another feature extraction required and an additional layer that will find relations to segments of firms made as following explanation is done.

$$s_{i,t}^{spc} = (\sum_{k=1}^{k} \log_n(n_{k,i,t}) d_{ij} | \sum_{k=1}^{k} \log_n(n_{i,t}+1))$$

B. Relationship with ESG performance

With large quantity and divergence in ESG activities there will be improvement in ESG performance of a firm/ organisation. Thus, we concludes that there is positive relation between out scores and ESG performance of firm. We cross verified our hypothesis by comparing these score data against Thomson Reuters Asset4 score. KPI's are the generalised scores ranging from 0 to 100, also our score range is between 0 to 100 only while industrial average bounds upto 50. Scaling and scoring sometimes cannot fit against empty segments of missing data so we used Tokyo Stock Exchange logarithm as control variable. over extent of year *t*. Regression as follows:

TABLE IV
INDICATOR SCORE

Score	No. of indicators	Example
Environment	73	Pollutant released Metric Tons of CO_2
Social	42	Women to Men ratio
Governance	56	Anti-slavery policy Board meetings CEO compensation

Tokyo logarithms exchanges occurred from 2006 to 2012. So adding control variable we get:

$$s_{i,t}^{qnt} = \sum_{k=1}^{k} \log_n(n_{k,i,t}+1) + \sum_{t=2006}^{2012} \gamma D_t$$

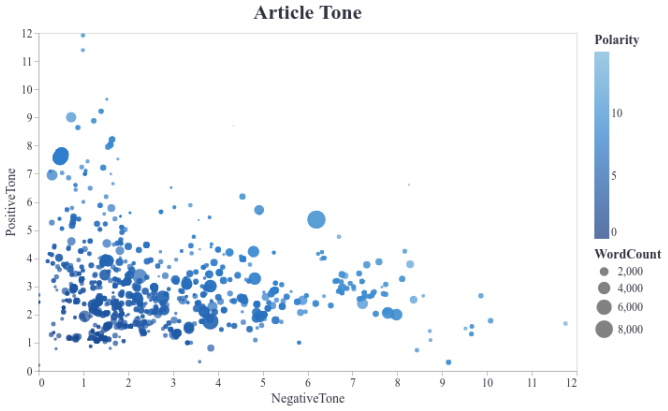

Fig. 4. Polarity - Word frequency

We confirm that Industrial average score are better parameter for analysing a ESG standard of firm as we go for higher variables i.e, more number of investment, sub divisions in target firm/organisation. Moreover, calculating industrial average of each topic environmental activity, social related activity and governance related activity there's an improvement of ESG quantitative performance from CSR reports.

C. Comparative ESG strategy

Over secondary objective in this proposed model is to provide an better investment option while we analyse our target firm/organisation. For example primarily we have chosen firm 'A' to invest in and we analyse it's ESG activities traditional analyst will be targeting that firms data and make it's decision if that investment is good or bad. However, in such case simple use of Node2vec specifically word2vec can ease our findings for an alternative option to investment in and again whole analysis part will be covered by the model.

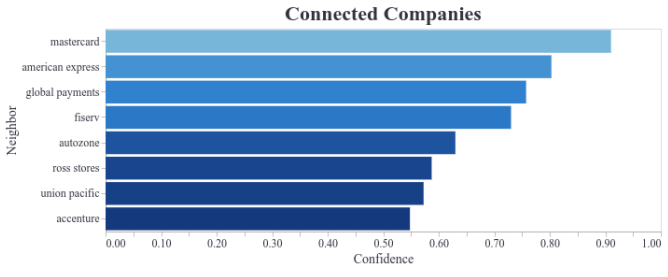

Fig. 5. Related firm confidence plots

Word2vec implements uni-gram natural algorithm to find the phrases from whole bucket data set create a distributional hypothesise to find the alternative investment, we used word

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 03, March-2023

embedding out to plot a 3D graph which will be visual output for easier accessibility to the feature. Term frequency is confidence factor in above Figure 5. here that is calculated by using document inverse frequency manner:
$W_{i,j} = tf_{i,j} * log(n|df_i)$

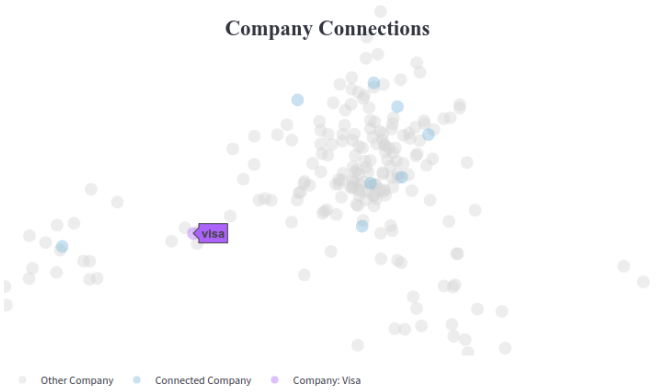

Fig. 6. Final embedded 3D graph

V. Conclusion

Upcoming years ESG will be a fundamental in sustainable growth and factor all investments in market. The proposed paper analysed text data of different firms and organisation over subject 'ESG activity' specifically. Traditional method of manual analysis is time consuming and found to be somewhat inaccurate as proper time stamping and real time data not available all time. This calls for mining model to visualise firm's/organisation's ESG activities by acquiring text documented data published according to global standards. We tested it over 100+ organisation and hypothesis of evaluating ESG scores by automation of natural language processing and various machine learning techniques proves to be better and faster solution against disadvantaged factors of manual work and greenwashing subject by organisations. Figure and

TABLE V
ESG PERFORMANCE

Category	ESG Performance		
	E	***S***	***G***
Adjusted R^2	0.224	0.233	0.296
E score	3.8*		
Specific E score	0.6608***		
Average ind. E score	11.21**		
S score		19.46*	
Specific S score		0.417***	
Average ind. S score		27.45**	
G score			10.45*
Specific G score			0.433***
Average ind. G score			19.42**

*Standard errors: *p¡0.05; **p¡0.01; ***p¡0.001*

Table is summary of subsequent work done in this paper. We found aspects of ESG performance from firm's CSR report(firm name not disclosed) and can be concluded that further investigation and proper sustainable methods could be taken to improve ESG ratings.

VI. References

[1] Azhar, Nurul Asyikeen, et al. "Text analytics approach to examining corporate social responsibility." Asian Journal of Accounting and Governance 11 (2019): 85-96.
[2] Mack, Daniel, et al. "Reaching students with Facebook: Data and best practices." (2007).
[3] Lamba, Manika, and Margam Madhusudhan. "Text Mining for Information Professionals."
[4] Patra, Swapan Kumar. "How Indian libraries tweet? Word frequency and sentiment analysis of library tweets." (2019).
[5] Friedrich, Natalie, et al. "Adapting sentiment analysis for tweets linking to scientific papers." arXiv preprint arXiv:1507.01967 (2015).
[6] Székely, Nadine, and Jan Vom Brocke. "What can we learn from corporate sustainability reporting? Deriving propositions for research and practice from over 9,500 corporate sustainability reports published between 1999 and 2015 using topic modelling technique." PloS one 12.4 (2017): e0174807.
[7] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).
[8] Levytska, S. O., et al. "ANALYSIS OF BUSINESS ENTITIES'FINANCIAL AND OPERATIONAL PERFORMANCE UNDER SUSTAINABLE DEVELOPMENT."(2018): 122-127.
[9] Kiriu, Takuya, and Masatoshi Nozaki. "A text mining model to evaluate firms' ESG activities: an application for Japanese firms." Asia-Pacific Financial Markets 27.4 (2020): 621-632.
[10] Ioannou, Ioannis, and George Serafeim. "The consequences of mandatory corporate sustainability reporting." Harvard Business School research working paper 11-100 (2017).
[11] Serafeim, George. "The Consequences of Mandatory Corporate Sustainability Reporting, Evidence from Four Countries." (2014).
[12] Keith Norambuena, Brian, Exequiel Fuentes Lettura, and Claudio Meneses Villegas. "Sentiment analysis and opinion mining applied to scientific paper reviews." Intelligent data analysis 23.1 (2019): 191-214.
[13] Baccianella, Stefano, Andrea Esuli, and Fabrizio Sebastiani. "Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining." Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). 2010.
[14] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques." arXiv preprint cs/0205070 (2002).
[15] Boiy, Erik, and Marie-Francine Moens. "A machine learning approach to sentiment analysis in multilingual Web texts." Information retrieval 12.5 (2009): 526-558.
[16] Sinha, Pritish, et al. "Education and Analysis of Autistic Patients Using Machine Learning." 2022 International Conference on Emerging Smart Computing and Informatics (ESCI). IEEE, 2022.