ML Project

Tuhin Harit & Serena Xiao

May 2021

1 Introduction

A recent article in WSJ "Green Euphoria may cost investors, but Planet says Thank you" reported that there is considerable amount of capital $80 billion followed into environmental, social and governance i.e. ESG funds in the third quarter of 2020. These funds only invest in companies that pass the stringent tests on meeting the ESG criteria – a group of standards used by socially conscious investors to screen investments.

Bubble in stock prices of the "green" companies has been a long concern for investors in this space. A fund tracking Nasdaq clean energy index had risen 191% between end of 2019 to early 2021, which is more than ten times higher than the gain from the broad market index. Quoted from this article: "Private markets generally provide too little incentive for risky innovation because shareholders only capture a small part of an innovation's benefit; most goes to consumers (think of a life-saving drug)". For the past decades, there have been increasing number of empirical studies exploring the relationship between firm performances and green innovation. However, the results are mixed, and it remains ambiguous how or whether the adoption of green innovation would affect a firm's performance. One obstacle that the researchers are facing in this area is the definition and measure of "green innovation".

Our main contribution to the literature is to adapt a relatively newer approach to measure "green innovation". Following the approach described in Li et al 2020, we adopt their technique to quantify text by using the word embedding model (Mikolov et al, 2013, aka word2vec). Based on the seed words and the texts extracted from the earnings call transcripts, we are able to construct a "green innovation dictionary". Then based on the frequency count of these words and phrases in the earnings call transcripts, we can generate scores for each firm-quarter. We contrast our measure with Green innovation measure constructed from Loughran and McDonald, 2011 (L&M) which uses a hand-collected dictionary seeded by "most frequently used words" from an authoritative literature on green innovation.

Our proposed measure using the machine learning approach provides a comprehensive scoring system. On one hand, this score considers the existing green product and process innovation within a firm, which are the two main categories have been studied in the existing papers. On the other hand, it captures the potential of future green development in the firm. We find that our measure of green innovation is significantly able to explain the "actual green innovation" by firms.

2 Related literature

Most of the existing research has either used a specific proxy for environmental performance of the firm or considered green innovation in two separate categories: green product v.s. process innovation. For example, Lee et al. 2015 studied the impact of CO2 emissions on firm performance. Driessen et al. 2013 and Albino et al. 2012 examine the green product innovation, and Tseng et al 2013 focuses specifically on the green process innovation. Focusing on one specific area of green innovation is sometimes problematic, as it is difficult to disentangle the effect from green product innovation, green process innovation and overall green policy adopted by the company. When green innovation exists in multiple areas within a firm, there could be spill-over effect. That motivates us to propose one single measure to measure the overall exposure of the firm to "green technology".

As mentioned previously, Loughran and McDonald (2011) use a hand-collected dictionary and use it to predict innovation measure. We closely follow the method proposed by Li et al (2020) for measuring innovation and extend it on green innovation and find that it works well in capturing innovation of firms with previous ESG activism (green firms).

3 Data

We select the companies with past ESG shareholder proposals. The sample period is 2005-2020. Our earnings call transcripts data are from WRDS database. We have included 14500 transcripts in our sample. The alternate (older) green innovation measure by L&M that needs hand-created dictionary seeded with most frequently occuring words derived from authoritative texts including research paper(E.g. The Driver of Green Innovation and Green Image: Green Core Competence by Chen et al 2008) and books (Assessment of Carbon Footprint in Different Industrial Sectors by Muthu).

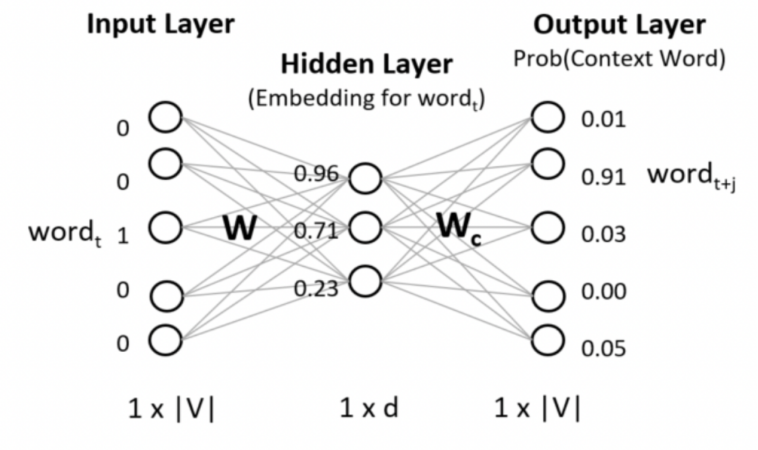
Our algorithm counts the number of mentions of our generated "green innovation" dictionary words in a transcript (Term Frequency - TF). We also generate TF-IDF which evaluates a word's relevance to a document in a collection of documents. (This is done by multiplying two metrics: how many times a word appears in a

document, and the inverse document frequency of the word across a set of documents). We also do WF-IDF which uses the log of term frequency in the numerator.

4 Methodology

We want to see how well our measure of green innovation (adapted from Li et al(2020)), performs in "actual" green innovation. In absence of a ready-made measure of "actual" green innovation, we use the R&D expenditure by green firms as the proxy for actual green innovation. We define green firms as those which have ESG activism in the past or future.

We follow the procedure adopted in Li et al. (2020) by applying the the Word2vec model introduced in Mikolov et al (2013) to calculate the word embedding vectors. Word2vec is a technique for natural language processing, which uses a neural network model to learn the word associations. It is an algorithm that takes text corpus as an input and outputs a vector representation for a given word. It utilizes neural network to train itself by reading through the texts in the earnings call transcripts. The training objective is to have the minimum prediction error across all context words in the outer layer.



source: Measuring Corporate Culture Using Machine Learning, Li et al 2018

The ultimate goal of the model is to predict neighboring words given an word input. It is equivalent to maximize the average log probability:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq c, j\neq 0} log(w_{t+j}|w_t)$$

c is the size of the training context, larger c can result in higher accuracy with larger amount of training time

Context window decides the range in terms of number of words that will be included as context words. We consider a word as "neighbor" if it is within five words from a given word. Choosing higher dimension typically

increases quality of word embedding, however the marginal gain diminishes. In our project, we choose our dimension to be 100. That means a word embedding coverts to a 100-dimensional vector that represents the meaning of the word. We use cosine similarity to quantify the distance between two words:

$$cosine(A, B) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{100} A_i B_i}{\sqrt{\sum_{i=1}^{100} A_i^2}\sqrt{\sum_{i=1}^{100} B_i^2}}$$

where A_i and B_i are components of vector A and B. The vectors are chosen such that the cosine similarity between the vectors measures the level of semantic similarity between the words represented by these vectors. After cosine similarity is calculated between any sets of words, we apply the commonly used method in information retrieval literature - bootstrapping to construct the green innovation dictionary.

After the dictionary is generated, we apply three different methods of calculating the score for each transcript: raw count, TF-IDF and TF-IDF with log normalization (WF-IDF). The raw scored is the frequency of dictionary words normalized by the total number of words. TF-IDF is a scoring measure that has the advantage to evaluate how important a word is to a document in a collection of documents. This measure is calculated by multiplying word frequency in a document with the inverse document frequency of a word across a set of documents.

As discussed previously, the alternate L&M measure requires hand-picked dictionary seeded from most frequently occuring words from representative text on green innovation topic. Once the seed-words are obtained, "hand-picking" is needed to cover the different grammatical forms of the sentiment within the same sentiment (for e.g. if technology is a green innovation seed-word, then technological (adj) and technologically (adverb) should also be "hand-picked" and included in green-innovation dictionary).

5 Results and discussion

Dictionary comparison with althernative method

As mentioned in methodology section, in absence of a direct measure we use r&d in green firms, which we define as firms that have atleast one occurrence of ESG motions, as a proxy of "actual" green innovation. Additionally, to highlight the superiority of our method over previous method used in finance literature (primarily L&M) we generate green innovation score using L&M.

Table below shows the top 20 words in the dictionary generated by these two methods. Out of these top 20 words, only 7 of them have appeared in both dictionaries. There are only 10 common words that appear in the top 60 words in both dictionaries (16.7%). We consider the word embedding model is a more advanced technique in our context, as it has the ability to learn the meaning of the words used in the transcripts.

It is more of a predictive model, whereas the alternative (LM) method is a count-based model. Dictionary generated by LM method involves many general words that are not closely tied to green innovation, such as "value", "role", "case", etc. In contrast, words dictionary generated by word2vec are more topically related.

Word2vec	innovation	energy_efficiency	technology	technological	environmental	energy	social	digitalization	energy_transition	electrification
	renewable_energy	innovative	development	sustainability	health_care	energy_infrastructure	automation	capability	environmentally	healthcare
LM	innovation	sustainable	business	research	energy	policy	environmental	new	social	management
	development	technology	sustainability	change	production	journal	systems	models	process	system

Validation results

To evaluate whether our measure truly capture the R&D effort and overall company strategy related to green innovation, we validate our results by using R&D expenses in the Compustat database. Green innovation spending should be part of the R&D expenses incurred by any company. We normalize R&D expense by company's revenue. Table below presents the results of validation tests using both the word2vec and L&M method. We show that both the contemporaneous and one-quarter lagged green innovation scores based on word2vec method are significantly positively correlated with the R&D expenses. This positive relation remains significant after control for industry and year fixed effect. In comparison, the green innovation index generated using L&M method does not show any significant association with the R&D expenses. These results suggest that the green innovation score generated by this word embedding model works as expected.

Word2vec	R&D/Rev	R&D/Rev
Green_inn_TFIDF	0.05	0.01
t-stat	***4.44***	***2.51***
1 Qtr Lagged Green_inn_TFIDF	0.014	0.01
t-stat	***4.59***	***2.63***
L&M		
Green_innovation index	0.00125	-0.0006
t-stat	*0.71*	*-1.1*
Control	Year FE	Year FE, Industry FE

The results are equally significant for TF and WF-IDF.

6 Conclusion

To conclude, we are able to extend Li et al (2020) results to create a measure of green innovation which effectively captures the actual green innovation in firms. We find that our measure is more effective in capturing green innovation compared to the traditional L&M measure. We assign this to superior ability to Word2vec to capture nuanced meanings of words and phrases which the previous methods lacked.

For future work, we would like to test our measure against more direct green innovation proxy (such as green patents). We would like to increase our dataset to include all firms (even those without ESG motions). We would also like to include a more effective benchmark to test the robustness of our results. A natural extension of this research is to use this measure to predict company performance.

References

Albino, V., Balice, A., Dangelico, R. M. *Environmental strategies and green product development: an overview on sustainability- driven companies.* Business Strategy and the Environment, 18(2), 83-96, 2012.

Driessen, P. H., Hillebrand, B., Kok, R. A., Verhallen, T. M. *Green new product development: the pivotal role of product greenness.* IEEE Transactions on Engineering Management, 60(2), 315-326, 2013.

Horváthová, E. *Does environmental performance affect financial performance? A meta-analysis.* Ecological Economics 70(1), 52-59, 2010.

Green Euphoria May Cost Investors, but Planet Says Thank You
`https://www.wsj.com/articles/green-euphoria-may-cost-investors-but-planet-says-thank-you-11609949560`

Kai Li, Feng Mai, Rui Shen, Xinyan Yan. *Measuring Corporate Culture Using Machine Learning.* The Review of Financial Studies, 2020.

Ki-Hoon Lee, Byung Min, Keun-Hyo Yook. *The impacts of carbon (CO2) emissions and environmental research and development (R&D) investment on firm performance* International Journal of Production Economics, 167, 2015.

Mikolov, T., I. Sutskever, K. Chen, G.S. Corrado, and J. Dean. *Distributed representations of words and phrases and their compositionality.* Advances in Neural Information Processing Systems, 3111- 3119, 2013.