Green innovation measure using word2vec

Tuhin Harit and Serena Xiao

University of Texas at Dallas

Introduction

- large fund flow into environmental, social and governance, i.e. ESG funds
- Existing literature seems to lack a good "green-innovation" proxy measure which unambiguously captures green innovation leading to ambiguous firm performance prediction
- Existing sentiment analysis papers (before Li et al) typically used manually constructed dictionaries
- Using Word2vec (i et al 2020) we create a green innovation sentiment dictionary that learns from the training set
- Using dictionary we create Green innovation score and find that it significantly predicts the "actual" green innovation by firms

Existing literature

- increasing number of empirical studies exploring the relationship between firm performances and green innovation (Lee et al 2015, Driessan et al 2013, Albino et al 2012 etc)
- Results are mixed - remains ambiguous how or whether the adoption of green innovation would affect a firm's performance
- **obstacle**: lack of consistent definition and measure of **green innovation**
- our contribution: create a green innovation measure using word2vec

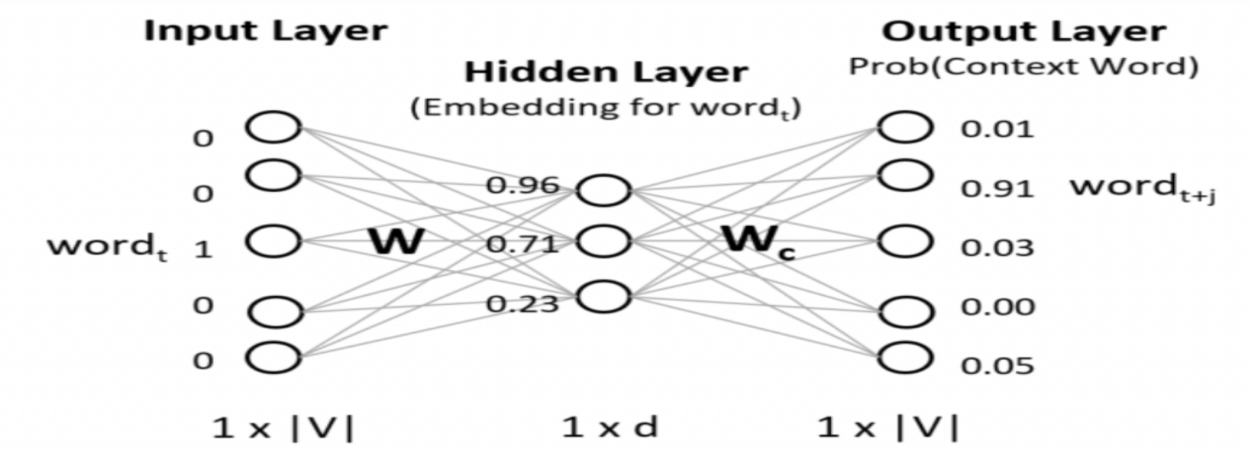
Data

Sample includes firms with past ESG shareholder proposals ("Green Firms").
R&D exp in these firms represents "True" green innovation.

- Review 14500 transcripts 450 firms covering period of 2005-2020
- Develop alternate L&M green innovation - hand-crafted dictionary seeded with most frequently occurring words derived from authoritative texts
- We do TF, TF-IDF (captures a word's relevance to a document in a collection of documents) and log TF-IDF

Methodology

- follow the procedure adopted in Li et al. (2020) to calculate the word embedding vectors and construct scoring system
- word2vec - technique for natural language processing, which uses a neural network model to learn the word associations
- ultimate goal of the model is to predict neighboring words given an word input



Methodology

- consider a word as neighbor if it is within five words from a given word
- a word embedding coverts to a 100-dimensional vector that represents the meaning of the word (dimension = 100)
- use cosine similarity to quantify distance between two words

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

Methodology

- use bootstrapping to construct **green innovation** dictionary
- calculate score for each transcript:
  - raw score: frequency of dictionary words normalized by the total number of words
  - TF-IDF: multiply word frequency in a document with the inverse document frequency of a word
  - TF-IDF with log normalization
- output - green innovation score for each transcript
- Alternate measure (L&M) generates seed words from authoritative research paper(E.g. Chen et al 2008) and books (e.g. Carbon Footprinting by Muthu)

Results - dictionary comparison with alternative method

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

- only 10 common words that appear in the top 60 words in both dictionaries (16.7%)
- L&M dictionary involves many general words that are not closely tied to green innovation, such as "value", "role", "case", etc

Results - validation

- evaluate whether the scores truly capture the RD effort and overall company strategy related to green innovation
- validate our results by using R&D expenses in the Compustat database
- both the contemporaneous and one-quarter lagged green innovation scores based on are **significantly positively** correlated with the R&D expenses
- The results are equally significant for TF and WF-IDF.

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

Conclusion

- We are able to extend Li et al (2020) results and create an effective green innovation
- Our measure is more effective in capturing green innovation than traditional L&M measure.
- In future, we would like to test our measure on more direct actual green innovation proxy (such as green patents).
- We would like to expand our dataset to include all firms
- A natural extension is to use this measure to predict company performance.