「A Study on the NLP Approach to Detect Greenwashing Signs of Companies in Korea」

2022. 10. 27

Yongjik Jay Lee

Contents

01 Intro.

02 Previous Studies on Data driven Greenwashing Detection

03 Our Approaches & Key Findings K-ClimateBERT V0.3

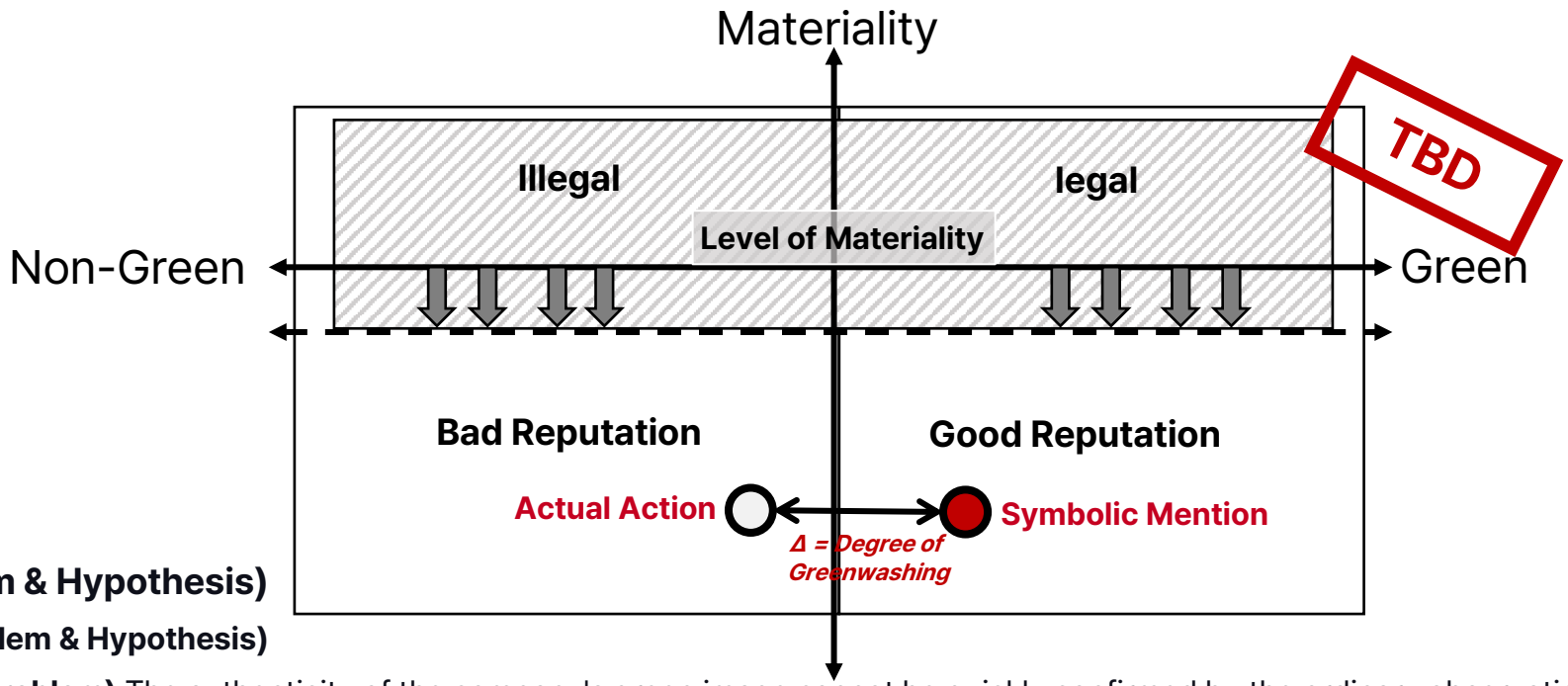
Overview of Our Research

- The recognition of 'Greenwashing' is not easy because there are limitations of 'information asymmetry' due to it being an undisclosed corporate’s internal information area's activities.
- This study aims to find a way to detect the authenticity of the company's green activities or the possibility of greenwashing through comprehensive and multi-faceted tracking of public information about corporate activities that anyone can easily acquire.
- To this end, considering the characteristics of greenwashing information composed of large-scale unstructured data, natural language processing (NLP) techniques based on big data and machine learning can be a meaningful solution.
- We are developing **the Korean Language based Climate Focused NLP Language Model (K-ClimateBERT)**, which can detect the authenticity of the company's green activities through tracking of public text information about corporate’s green activities
- Our Recent Ver.0.3 of K-ClimateBERT is targeting all 2,541 listed Korean corporates through NLP analysis of 0.7 million articles of media text information from the five major daily newspapers (Chosun, JoongAng, Donga, Hankyoreh, Kyunghyang), two economic newspapers (Hankyung, Meakyung), and five broadcasting media (KBS, MBC, SBS, OBS, YTN) for the past five years (2017-09-01 ~ 2022-09-01)

Key Hypothesis on Our Research

(Definition of Greenwashing)

Defined as a state of relative inconsistency between symbolic actions and actual actions related to the corporate's green image (identity)



(Problem & Hypothesis)

- **(Problem & Hypothesis)**
  - **(Problem)** The authenticity of the company's green image cannot be quickly confirmed by the ordinary observation of the general public, who do not have specific internal information about the company.
  - **(Hypothesis)**
    - By measuring the weight between the symbolic mention $^{\text{Cheap Talk, Implicit}}$ and actual action$^{\text{Explicit}}$ that companies use/expose daily in external communication media (news/media) and comparing it with the social standard on Green talk behavior of industries, the degree and possibility of Greenwashing can be measured.
    - In addition, by measuring the quantifiable greenwashing position of companies, it is expected that the practical effect of social activities of greenwashing monitoring can be increased by limiting/selecting subjects that require intensive observation of greenwashing.

ClimateBERT: A Pretrained Language Model for Climate-Related Text

Summary

By Nicolas Webersinke, Mathias Kraus(FAU Erlangen-Nuremberg, Germany), Julia Anna Bingler (ETH Zurich, Switzerland), Markus Leippold(University of Zurich, Switzerland)

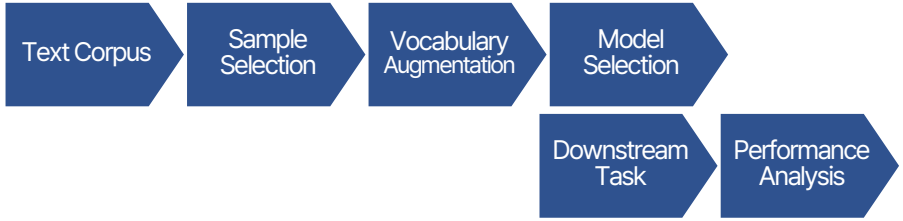
Subjects: **Computation and Language (cs.CL)**
Cite as: arXiv:2110.12010 **[cs.CL]**
or arXiv:2110.12010v1 **[cs.CL]**
https://doi.org/10.48550/arXiv.2110.12010
Submission history
From: Markus Leippold
[v1] Fri, 22 Oct 2021 18:47:34 UTC (3,545 KB)

Abstract

Over the recent years, large pretrained language models (LM) have revolutionized the field of natural language processing (NLP). However, while pretraining on general language has been shown to work very well for common language, it has been observed that niche language poses problems. In particular, climate-related texts include specific language that common LMs can not represent accurately. We argue that this shortcoming of today's LMs limits the applicability of modern NLP to the broad field of text processing of climate-related texts. As a remedy, we propose ClimateBert, a transformer-based language model that is further pretrained on over 1.6 million paragraphs of climate-related texts, crawled from various sources such as common news, research articles, and climate reporting of companies. We find that ClimateBert leads to a 46% improvement on a masked language model objective which, in turn, leads to lowering error rates by 3.57% to 35.71% for various climate-related downstream tasks like text classification, sentiment analysis, and fact-checking.

Approaches

- Many studies based on traditional NLP approaches in Green Monitoring face considerable limitations, since climate related wording could vary substantially by source (Kim and Kang, 2018).
- Deep learning techniques promise higher accuracy. (e.g., Kölbel et al., 2020; Luccioni et al., 2020; Bingler et al., 2021; Callaghan et al., 2021; Wang et al., 2021; Varini et al., 2020).
- One of the most prominent NLP models is called BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018)



① **Domain Specific pre-training** : As climate change is a complex, fast-moving, and often ambiguous topic with scarce resources, extracting climate-related information from textual sources is challenging

- **Climate Text Corpus :** Pretrainned 1,662,206 climate-related paragraphs of text from news articles, research abstracts, and corporate climate reports

② **Downstream Task :** Text classification, Sentiment analysis, Fact-checking → Cheap talk in corporate climate commitments: The role of active institutional ownership, signaling, materiality, and sentiment (Nicolas et al., Jan. 2022)

Step 1: Segmentation & Sampling

	KOSPI	Kosdaq	Konex	SUM
Fossil fuel, Chemicals and material products	141	222	21	**384**
Electronics and Precision instrument manufacturing	46	300	16	**362**
Manufacture of electrical, mechanical and other products	62	260	17	**339**
Information/Communications/Broadcasting and Media Services	34	237	21	**292**
Finance and Insurance	105	110	-	**215**
Logistics, Distribution and Transportation Industry	92	100	5	**197**
Manufacturing of automobiles and transportation equipment	61	67	5	**133**
Manufacturing of iron and metal products	53	48	2	**103**
Construction and Real estate	58	39	5	**102**
Education, R&D and Human Service	27	98	23	**148**
Textile and consumer goods manufacturing	38	29	4	**71**
Food and Beverage Manufacturing	26	32	2	**60**
Mining, mineral products and cement manufacturing	24	14	1	**39**
Wood product manufacturing	24	12	1	**37**
Natural resource production and processing	15	11	2	**28**
Environmental & Energy Services	12	9	-	**21**
Accommodation and leisure industry	5	5	-	**10**
	823	**1,593**	**125**	**2,541**

Approaches & Implication

At K-ClimateBERT v0.1, out of 2,499 listed companies in Korea, 549, about 20% + α (2%), were derived.

Our latest research for K-ClimateBERT V0.3 includes all of the listed companies in Korea.

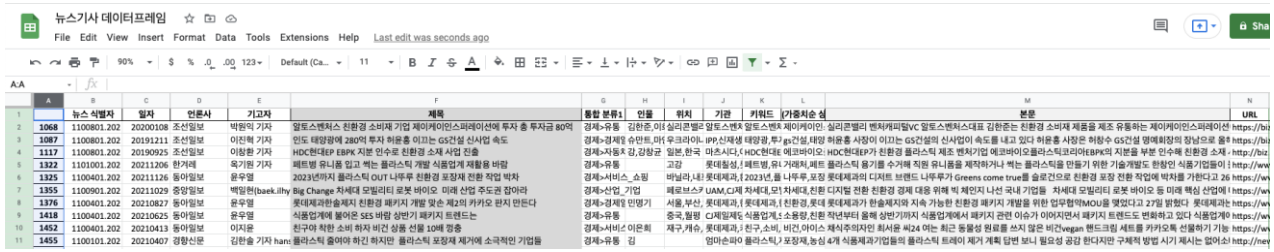
195 out of 821 KOSPI
334 of 1,549 KOSDAQ
20 of KONEX 129

20% for each industry group, 20% + α ~ all sampling if the parameters are below the average for analysis of the characteristics of each industry group

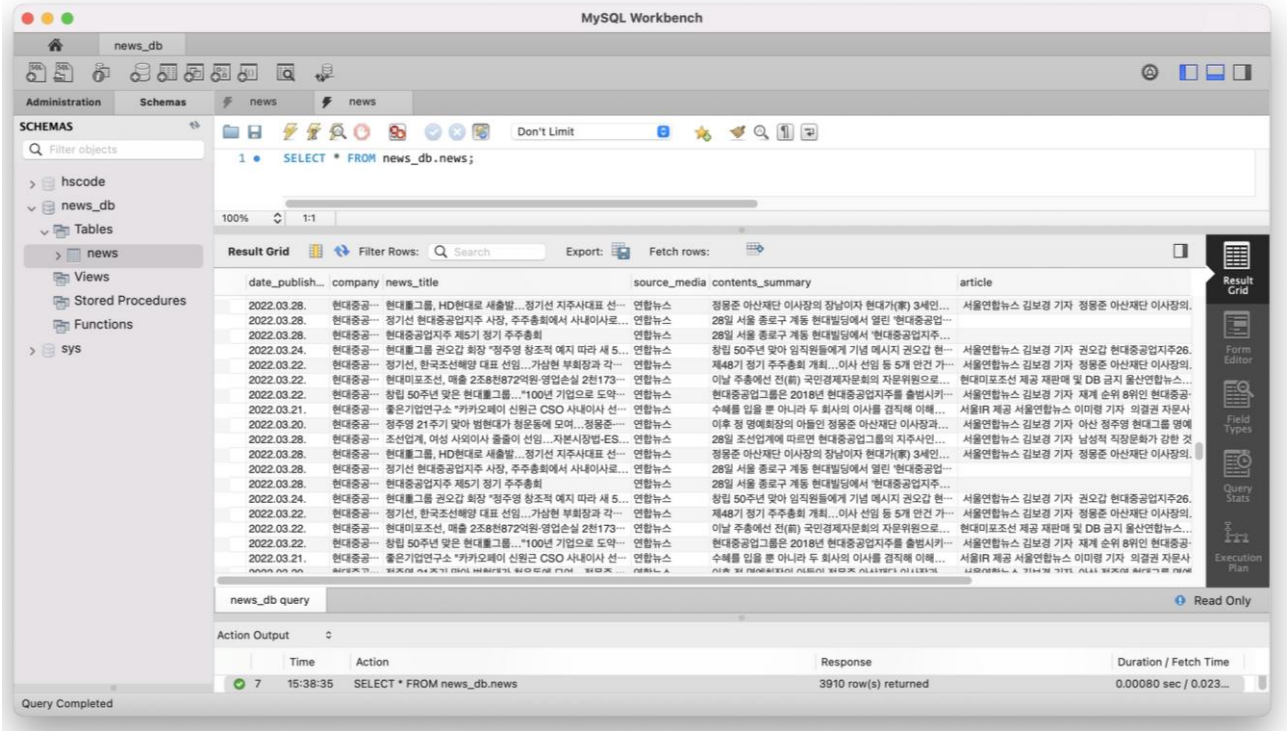
823 KOSPI
1,593 KOSDAQ **2,541**
125 KONEX

K-ClimateBERT V0.3

Step 2: Crawling of news media information by company & pre-processing of data



Database of news articles generated by crawling code (MySQL)



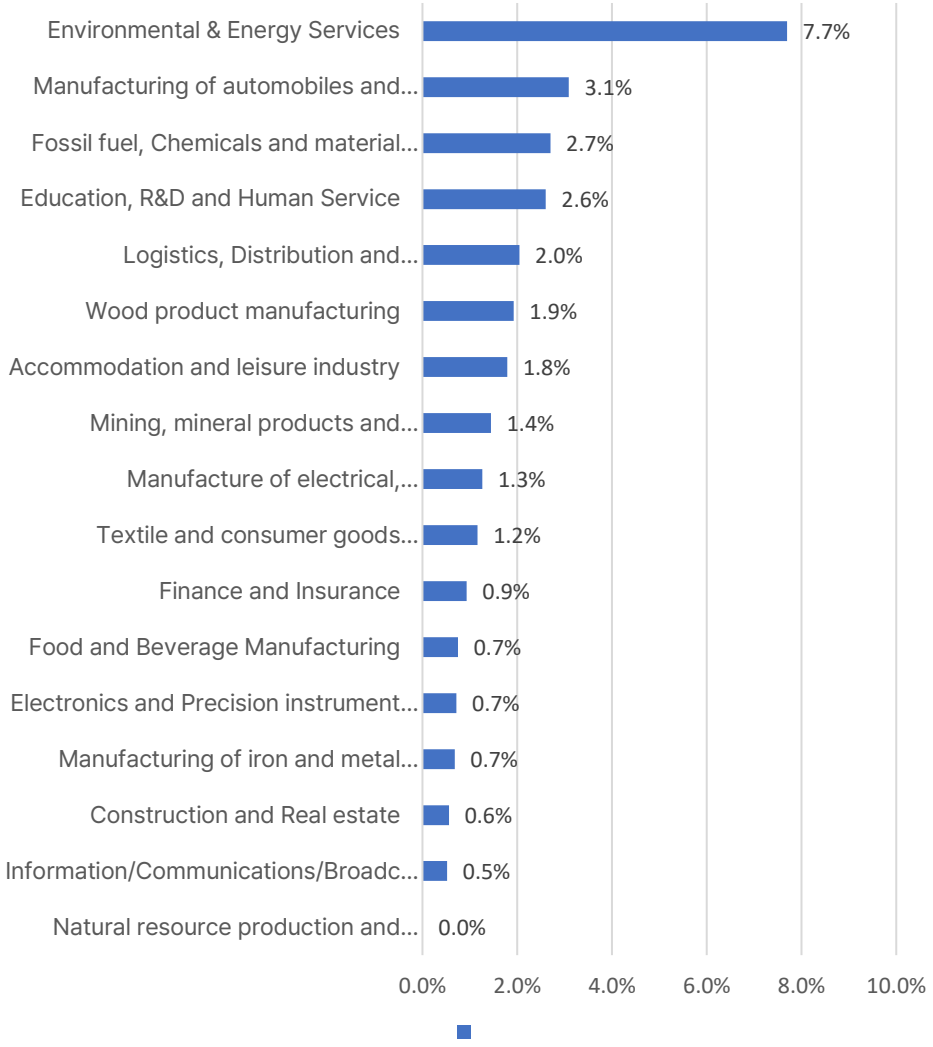
Approaches & Implication

- Crawling News media exposure information and data collection
- **Data Source: Online news articles from the five major daily newspapers (Joseon, JoongAng, Donga, Hankyoreh, Kyunghyang), two daily economic newspapers (한경, 매경), and five broadcasting media KBS, MBC, SBS, OBS, YTN for the past five years (2017-09-01 ~ 2022-09-01)**
- **Method:** Using the Korea Press Foundation's Big Keys API, extract a list of news articles on selected sample companies, and analyze the metadata (media company, contributor, title, etc.) and entity name (person, institution, place, etc.) of the newsText data
- **preprocessing**
  - Hangul Natural Language Processing (NLP) Precedent: Converting Input Strings to Morphological Columns

    * Morphological analysis refers to the process of separating words, which are words with spaces, into morphemes, which are the minor units of meaning that can no longer be analyzed.
  - This study deleted stopwords such as special characters and punctuation marks from article data, and tokenization was performed.
  - To further increase analysis accuracy, only significant parts of speech, such as nouns, verbs, adjectives, and adverbs, are extracted and applied.
  - Bot creation unrelated to Green Talk among news articles Article articles in the form of stock market conditions (e.g., Chosun Ilbo C-Biz bot), duplicate reports, and personnel transfer/obituary-related messages are excluded through exception handling.

Step 3 : GreenTalk Identification by K-ClimateBERT and Relation Analysis Between GreenTalk Behavior & Scientific Results by Industries

GreenTalk Ratio by Industry



Key Findings & Implication

Rank	Industry	GreenTalk Ratio	CO2 Reduction Rate*
1/17	Environmental & Energy Services	7.69%	-11.2%
2/17	Manufacturing of automobiles and transportation equipment	3.09%	-2.1%
3/17	Fossil fuel, Chemicals and material products	2.70%	-2.0%
5/17	Logistics, Distribution and Transportation Industry	2.04%	-5.3%
6/17	Wood product manufacturing	1.92%	-1.9%
	.....		
15/17	Construction and Real estate	0.56%	2.5%
16/17	Information/Communications/Broad casting and Media Services	0.52%	2.1%

* CO2 Reduction Rate (2017 ~ 2020, CAGR)

More GreenTalker make Better Performance on climate contribution !

Step 4 : Green Talk Classification ($N_{Explicit}$ / $N_{explicit+Implicit}$)

①Main-body Part ②Action Part

SK Innovation held a seminar 'Double Botton Line Insight Week' to share eco-friendly business models with social ventures.

에스케이(SK)이노베이션이 소셜벤처와 머리를 맞대고 친환경 비즈니스 모델을 공유하는 세미나 '디비엘(Double Botton Line) 인사이트 위크'를 개최했다

Main-body Part + Action Part

→ Implicit Mention

①Main-body Part ②Action Part ③Result Part

LG Chem and LG Energy Solutions, a battery subsidiary, recently invested 60 billion won in Li-Cycle, the largest battery recycling company in North America.

LG화학과 배터리 자회사인 LG에너지솔루션은 최근 북미 최대 배터리 재활용 업체 '라이사이클(Li-Cycle)'에 600억원 규모의 지분 투자를 했다

Main-body Part + Action Part + Result Part

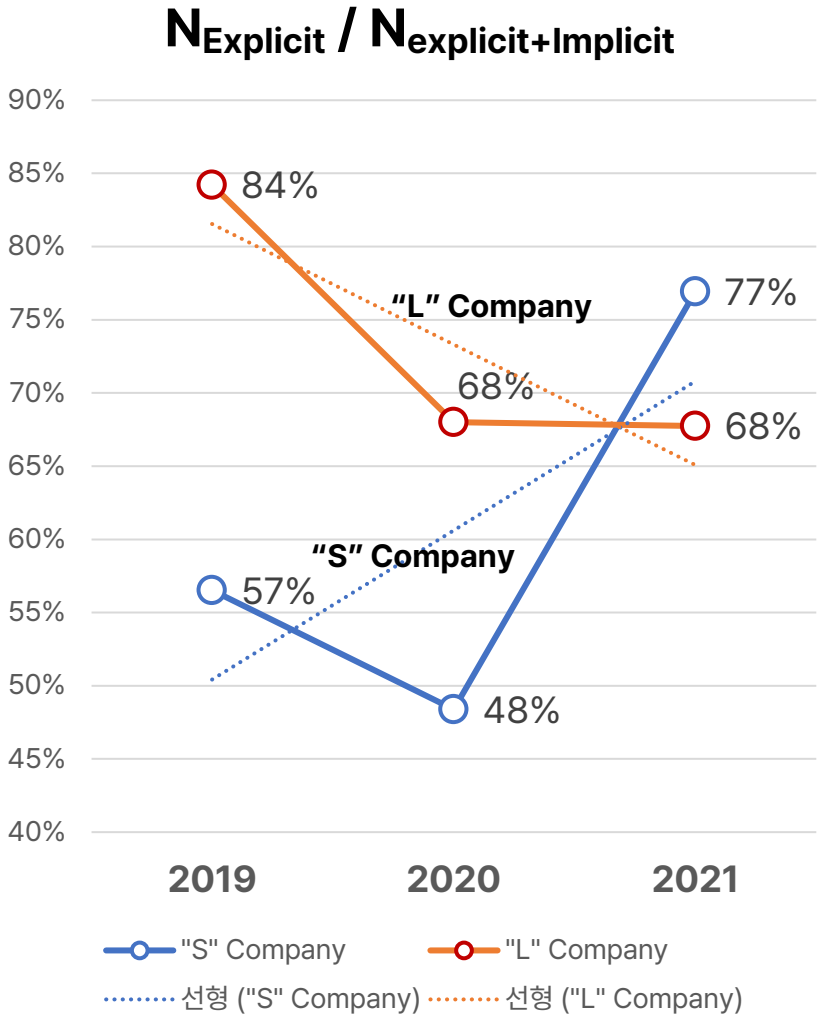
→ Explicit Mention

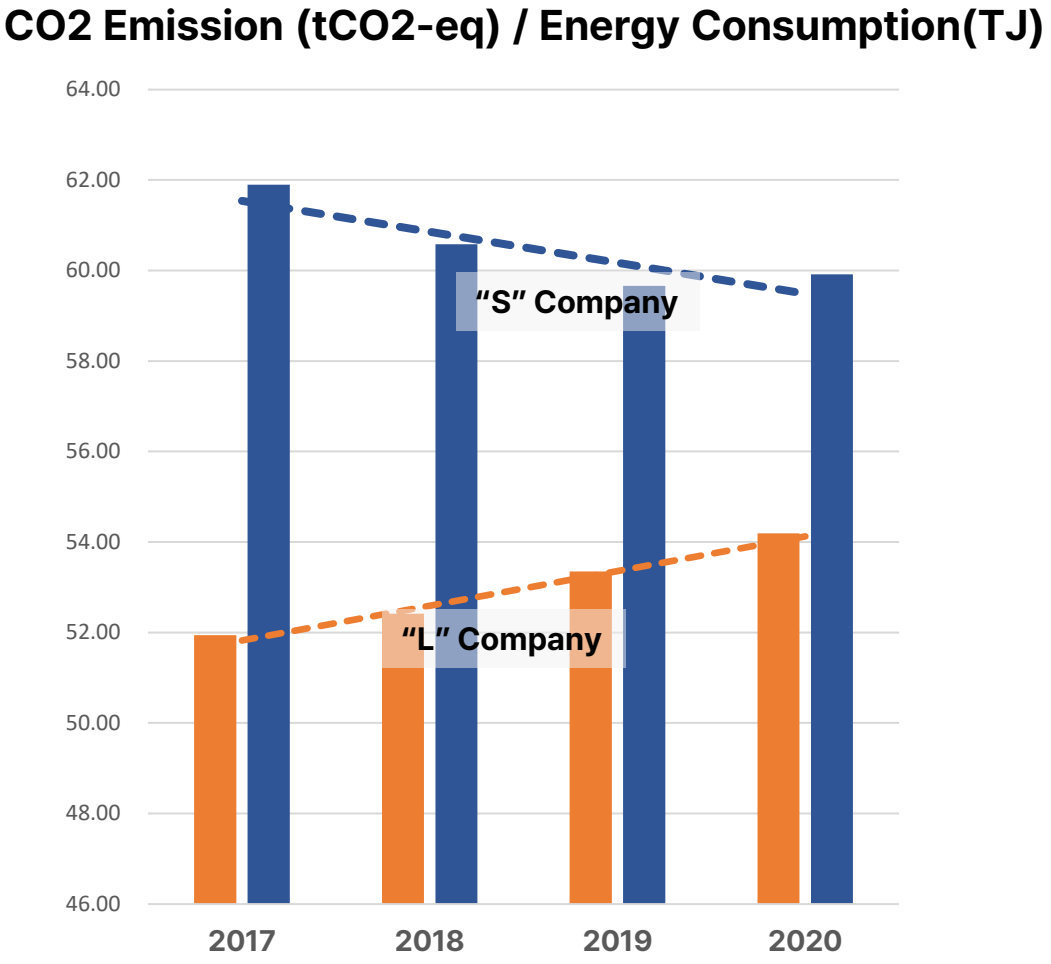
Does Better Quality of GreenTalker make Better Performance of Contributing to the climate?

Performance Analysis

“S” Company vs. “L” Company

(Fossil fuel, Chemicals and material products Manufacturing






YONSEI UNIVERSITY
1885
연 세 대 학 교