Transformer 기반의 토픽 모델링을 이용한 지속가능경영보고서 분석

이한울 [1], 이지현 [2], 이준희 [2]
[1] 고려대학교 사회학과
[2] 고려대학교 통계학과
gksdnf424@korea.ac.kr, qkffpsxkdlwm@korea.ac.kr, ljhee888@naver.com

Sustainability Report Analysis Using Transformer-Based Topic Modeling

Hanwool Lee[1], Jihyun Lee[2], Junheui Lee[2]
[1]Dept. of Sociology, Korea University
[2]Dept. of Statistics, Korea University

요 약

기업의 사회적 책임에 대한 요구가 높아짐에 따라 기업의 지속 가능 경영 보고서 발간은 증가 추세를 보이고 있다. 그러나 이전까지의 연구는 지속가능성 및 기업의 재무적, 비재무적 연관성에 초점이 맞춰져 있었으며, 전통적인 토픽 모델링 기법만을 제한적으로 사용한다는 한계를 보였다. 본 연구에서는 Transformer 기반의 맥락을 고려한 토픽 모델링 기법을 도입하여 다양한 이해관계자 측면에서 이용 가능한 25 개의 주제를 도출하였다. 또한 동적 토픽 모델링(Dynamic Topic Modeling)을 통해 주제의 변화를 시계열적으로 파악했다.

1. 서론

최근 기업의 사회적 책임에 대한 요구가 높아짐에 따라 기업의 지속 가능 경영에 대한 논의가 활발히 이루어지고 있다. 이에 기업들의 지속가능경영보고서 발간이 증가 추세를 보인다. [1].

지속가능경영보고서란 기업이 지속가능경영에 대한 전략의 방향성 및 비재무적 정보의 성과를 정리한 보고서다. 이는 기업 이해관계자의 소통 수단, ESG 평가 자료 등으로 활용된다. 또한 점차적으로 지속 가능한 회계 기준, ESG 공시가 단계적으로 의무화 됨에 따라 기업의 지속가능경영보고서에 대한 중요성이 대두되고 있다.

그러나 기업의 지속가능경영에 대한 기존 연구들은 주로 지속가능성과 기업의 재무적, 비재무적 연관성을 파악하는 데 집중되어 있다[2]. 또한 지속가능경영보고서에 관한 토픽 모델링 연구도 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA)과 같은 전통적인 토픽 모델링 기법을 사용한다는 한계를 보인다.

따라서 본 연구에서는 10 년간의 기업 지속 가능 경영 보고서에 대한 토픽 모델링을 통해 지속 가능성 이슈를 심층적으로 파악하고자 한다. 또한 전통적인 토픽 모델링 기법에서 벗어나 Transformer 기반의 모델을 활용해 맥락 파악이 가능한 토픽 모델링을 한다는 의의를 가진다.

2. 이론적 배경

토픽 모델링은 문서 군집으로부터 주제를 발굴해내는 기법을 의미한다.[3] 잠재 디리클레 할당과 같은 전통적인 토픽 모델링 기법들은 다양한 분야에서 활용되어왔으며 그 유용성을 입증해왔다.

최근 몇 년간 인공신경망을 활용하는 토픽 모델링 기법들이 높은 성과를 보이고 있다.[4] 전통적인 모델에 인공 신경망 기반의 단어 임베딩을 결합하는 것은 BoW(Bag of Word)에 의존했을 때 나타나는 다양한 한계점을 개선해 주는 것으로 알려져 있다.[5]

특히 Transformer 를 기반으로한 언어 모델들은 기존의 RNN 을 이용한 언어모델보다 다양한 벤치마크에서 높은 성과를 보였다.[6] 이를 바탕으로 Transformer 기반의 사전 학습된 모델을 활용하여 문맥을 반영한 토픽 모델링을 시도한 경우도 다수 존재한다. 그중 대표적인 토픽 모델링 기법인 복합 토픽 모델링(CTM, Combined Topic Modeling)은 BERT 와

BoW 를 결합하는 방식으로 토픽 모델링을 수행한다.[5]

BERTopic 은 이전의 복합 토픽 모델링과는 다르게 BoW 의 요소를 제거하고 사전 학습된 언어 모델을 적극적으로 활용하였다. BERTopic 은 다양한 데이터에서 기존의 다른 토픽 모델링 기법의 성능을 능가하는 것으로 나타났다.[7]

본 연구에서는 BERTopic 방법론을 이용하여 지속가능경영 보고서에 대한 토픽 모델링을 진행하고자 한다.

3. 모델 개발

3.1 데이터

본 연구의 분석 대상은 대한민국지속가능성지수(KSA)의 지속가능성 보고서 DB 에 게시된 2011 년부터 2021 년까지의 상장사 및 비상장사 1,041 개의 지속가능경영보고서다. 지속가능성 보고서 DB 에 있는 pdf 보고서를 OCR 을 통해 텍스트 데이터로 변환하여, 그 중 CEO 메시지 및 경영진 메시지를 추출하였다. 이렇게 추출된 데이터는 전체 지속가능보고서 데이터의 13.3%에 해당한다.

3.2 실험 환경

GPU : Nvidia Tesla P-100 (16GB)
CPU: Intel® Xeon CPU 2.30GHz, 6 core
RAM: 53.483 GB
Python version: 3.7.13

3.3 모델링 과정

BERTopic 은 총 3 단계를 거쳐 토픽 모델링을 진행한다. 먼저, 사전 학습된 언어 모델을 이용하여 각 문서에 대한 임베딩 벡터를 추출한다. 이후 UMAP[8]을 이용하여 클러스터링을 위한 차원 축소를 실시한다.

두 번째로, 차원 축소를 한 임베딩 벡터에 대해 HDBSCAN 을 이용해서 클러스터링을 진행한다. 이 과정을 통해 문서들의 의미론적 클러스터들이 나타난다. 최종적으로는 이렇게 만들어진 클러스터에 대한 주제 표현(topic representation)을 도출하기 위한 계층화된 TF-IDF 를 수행하게 된다.

본 연구에서는 언어 모델로서 Sentence-BERT 를 이용하였다. Sentence-BERT 는 Transformer 기반 모델로서 코사인 유사도를 이용할 수 있는 문장 임베딩 벡터를 도출해 주는 모델이다.[9] 본 연구에서는 해당 모델을 한국어로 이용하기 위해 100 가지 언어에 적용 가능한 huggingface 의'xlm-r-100langs-bert-base-nli-stsb-mean-tokens' 사전 학습 언어 모델을 활용하였다. [9]

주제 표현을 도출하는 과정에서는 각 토픽의 특징을 잘 파악하기 위하여 다양한 품사를 포함할 수 있는 RhinoMorph 형태소 분석기를 이용하였다. NNG(일반 명사), NNP(고유 명사), NP(대명사), VV(동사), VA(형용사), XR(어근), MM(관형사), MAG(일반 부사), MAJ 의(접속 부사) 품사를 활용하여 효과적인 주제 표현을 도출하고자 했다.

4. 결과

4.1 모델 결과

우리는 다양한 초모수 조정을 통하여 25 개의 주제를 지닌 최적의 모델을 도출하였다.

<표 1> 주제와 주제 표현

주제	주제 표현
주제 1	사업, 이사회, 위하다, 이슈, 강화
주제 2	에너지, 친환경, 환경, 온실가스, 기술
주제 3	윤리경영, 윤리, 리스크, 부패, 교육
주제 4	먹거리, 식품, 농산물, 바르다, 농식품
주제 5	기술, 사업, 성장, 혁신, 미래
주제 6	금융, 고객, 따뜻하다, 투자, 서비스
주제 7	공항, 항공, 세계, 여객, 서비스
주제 8	모바일, 플랫폼, 콘텐츠, 인터넷, 스마트폰
주제 9	중국, 아시아, 유럽, 글로벌, 사업
주제 10	아이, 미래, 아동, 청소년, 교육
주제 11	물관리, 수질, 국민, 관리, 수자원
주제 12	반도체, 메모리, 실리콘, 디스플레이, 제품
주제 13	대한민국, 건설, 미래, 발전, 철도
주제 14	화장품, 브랜드, 뷰티, 피부, 시장
주제 15	보험, 상품, 고객, 보장, 니즈
주제 16	스포츠, 국민, 건강하다, 건전, 즐기다
주제 17	화학, 소재, 차별화되다, 성장하다, 제품
주제 18	재활용, 폐기물, 원료, 생산하다, 국내
주제 19	백신, 신약, 제약, 의약품, 바이오
주제 20	여성, 돌보다, 지원, 성평등, 사회
주제 21	디지털, 전환, 혁신, 변화, 역량
주제 22	국제회계기준, 재무, 작성되다, 재무정보, 한국
주제 23	자원개발, 자원, 광산, 광물, 수급
주제 24	글로벌, 제품, 소재, 세계, 플랜트
주제 25	리스크, 관리, 재무, 유동성, 신용

<표 1>에서는 25 개의 주제와 각 주제에 해당하는 주제 표현이 나타나 있다. 주제 표현은 각 주제별 클러스터에 대한 TF-IDF 를 통해 나타난 빈출 표현들 중 다섯 개를 선정하였다.

4.2 결과 평가

토픽 모델링 기법의 평가 기준은 여러 가지가 있으나 대표적으로 주제 일관성(topic coherence)[10]과 주제 다양성(topic diversity)[11]이 활용된다. 주제 일관성은 주제 내의 단어들이 얼마나 의미론적으로 유사한지 확인할 수 있는 지표이고, 주제 다양성은 주제 간의 단어가 얼마나 상이한지에 대한 수치화를 통해 얼마나 다양한 토픽을 파악하고 있는지 확인할 수 있는 지표이다. 본 연구에서는 주제 일관성의 평가 지표로 gensim 의 topic coherence 지표[10]를 활용하였으며, 주제 다양성의 평가 지표로서 pairwise jaccard diversity 지표[11]를 활용하였다.

BERTopic 모델의 비교 대상으로는 이전 선행연구에서 활용되었던 잠재 디리클레 할당 모델을 사용하였다.

<표 2> 모델 평가

기법	주제 일관성	주제 다양성
LDA	0.6034	0.7901
BERTopic	0.6186	0.9652

<표 2>에서 BERTopic 모델은 주제 일관성과 주제 다양성 측면에서 잠재 디리클레 할당 모델을 능가하는 것으로 나타났다.

4.3 결과 분석

지속가능경영에 대한 사례 분석을 한 선행 연구에서는 국내 기업의 이해관계자별 중점 이슈를 제시하였다. [12] 본 연구에서는 제시된 이해관계자별 중점 이슈에 따라 모델에서 도출된 주제들을 분류하였다.

<표 3> 이해 관계자별 중심 이슈와 주제 번호

이해 관계자	중심 이슈	주제 번호
내부 경영 시스템	지속 가능 경영 강화	1
	국제 회계 기준 준수	22
	재무 위험 관리	25
주주 및 투자자	운영 투명성 및 윤리 경영 제고	3
산업 현장	산업 폐기물 재활용 방안 구축	18
	차별화된 소재 개발	17,24
	디지털 전환 및 산업 혁신	5,8,21
지역사회	아동 및 청소년을 위한 지원 프로그램 방안	10
	사회 공헌 활동	16
	성평등 지원 정책	20
환경	온실가스 관리	2
	미래 지향적인 국가 산업 방향성 제고	13
	자원 개발 및 관리	11,23
기타	산업별 이슈	4,6,7,12,14,15,19

<표 3>에 이해관계자 별로 중심 이슈와 이슈에 해당하는 주제 번호가 나타나 있다.

여러 선행 연구에서 제시되었던 토픽 모델링의 모델들은 지역사회, 환경, 내부경영시스템에 해당하는 주제만을 도출했다.[13] 또한 주제 표현도 단어로 제한되었다. [2], [13]

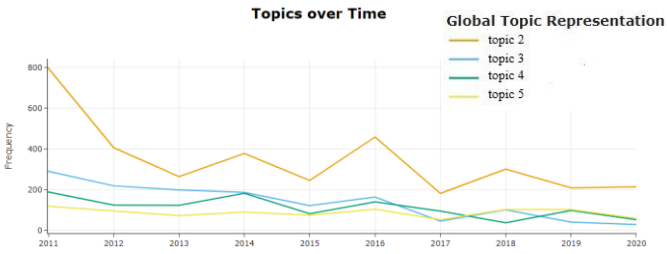
반면, 본 연구에서는 다양한 이해관계자와 주제들을 포착했다. 또한 <표 1>에서와 같이 주제 표현에 다양한 품사가 활용되어 주제를 식별하기에 용이했다.

4.4 동적 토픽 모델링

동적 토픽 모델링은 시간에 따라 유동적인 토픽 모델링 기법으로, 각 주제가 시간에 따라 등장하는 것을 추적할 수 있다. [14]

본 연구에서는 10 년간의 보고서에서 (모든 시간대에서) 가장 주요하게 나타나는 4 개의 주제를 활용하여 동적 토픽 모델링을 수행하였다. (정규화 삭제) 빈출 주제인 주제 2 – 주제 5 에 대해서 시간 변화에 따른 빈출도를 시각화 한 결과가 <그림 1>이다.

<그림 1> 동적 토픽 모델링



가장 많이 등장하는 주제 2 는 에너지, 친환경 기술 등에 해당하는 주제로 시간이 흐름에 따라 점차 감소하는 경향을 보이고 있다. 하지만 4 개의 주제 중에서 빈도가 가장 낮은 주제 5 는 성장 기술 및 혁신을 뜻하며 시간에 걸쳐 증가하는 양상을 보이고 있다. 그리고 먹거리, 농산물 등에 해당하는 주제 4 는 2014 년에 상대적으로 자주 등장하는 경향을 보였으나 그 이후에 지속적으로 감소하고 있다.

BERTopic 모델을 이용한 동적 토픽 모델링을 통해 시간에 따른 ESG 주요 주제의 변화 추이를 쉽게 파악할 수 있다. 기존의 연구에서는 연도별 특정 토픽의 등장 횟수를 기록하는 방식으로 ESG 트렌드를 확인했다. 반면에 동적 토픽 모델링을 이용해 ESG 특정 주제의 시계열 특징을 용이하게 파악하는 모습을 <그

림 1>을 통해 확인할 수 있다.

5. 결론

본 연구는 BERTopic 방법론을 활용함으로써 지속가능경영보고서에 대한 문맥을 반영한 토픽 모델링을 수행하였다. 이를 통하여 기존의 전통적인 방법론보다 더욱 다양하고 많은 주제를 검출해낼 수 있었고 주제 일관성 및 주제 다양성 측면에서도 우수한 결과를 나타냈다.

하지만, CSR 중심의 BoW 를 사용한 이전 연구들과는 달리 본 연구는 전적으로 Sentence-BERT 에 의존하였다. 그로 인하여 각 산업 군에서 나타나는 특수한 경우들도 각각의 주제로 검출되는 결과를 보였다.

또한, 의미론적 클러스터로부터 주제 표현을 도출하였기 때문에 각 주제 표현에서 키워드가 겹치는 현상도 일부 발생하였다.

그럼에도 불구하고 본 연구는 지속가능보고서 분석에 transformer 기반의 문맥을 반영한 토픽 모델링을 적용한 최초의 사례로서, 자연어 처리를 응용한 질적 연구에 새로운 전환점이 되었다고 판단된다.

또한 잠재 디리클레 할당 기법으로는 수행하기 어려웠던 동적 토픽 모델링도 일부 행함으로써, 시계열적인 토픽 관찰에 대한 가능성도 제시하였다.

참고문헌

[1] 최윤형, 이기호, 이상명. "지속가능경영보고서의 중요성 분석을 통해 바라본 지속가능경영 이슈와 10년의 변화." Korea Business Review, 26.1, 2022, 125-148.

[2] 윤지혜, 이종화. "토픽모델링을 활용한 CSR 키워드 트렌드 분석.", 인터넷전자상거래연구, 21.5, 2021, 73-91.

[3] David M Blei, Andrew Y Ng,Michel I. Jordan ‘Latent dirichlet allocation’, Journal of Machine Learning Research, 3, 993-1022, 2003

[4] Federico Bianchi, Silvia Terragni, Dirk Hovy, ‘Pre-training is a Hot topic:Contextualized Document Embeddings Improve Topic Coherence’, Proceedings of the 59th annual Meeting of the Association for Computational Linguistics, Bangkok of Thailand , 2021, 759-766

[5] Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, Elisabetta Fersini,’Cross-lingual contextualized topic models with zero-shot learning’ ,P roceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, Entirely Online, 2021, 1676-1683

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, ‘BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding’, arxiv eprints, arXiv:1810.04805, 2019

[7] Maarten Grootendorst, ‘BERTopic: Neural topic modeling with a class-based TF-IDF procedure’, arxiv eprints arXiv:2203.05794 , 2022

[8] McInnes, J.Heanly., J.Melville, ‘UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction’, arxiv eprints arXiv:1802.03426, 2018

[9] Nils Reimer, Iryna Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-networks’, arxiv eprints arXiv:1908.10084, 2019

[10] Michael Roder, Andreas Both, Alexander Hinnenburg, ‘Exploring the Space of Topic Coherence Measures’, Proceedings of the Eighth ACM international Conference on Web Search and Data Mining, New York of United States, 2015, 399-408

[11] Nam Khanh Tran, Sergej Zerr, Kerstin Bischoff, Claudia Niederée, Ralf Krestel, ‘Topic Cropping: Leveraging Latent Topics for the Analysis of Small Corpora”, International Conference on Theory and Practice of Digital Libraries,Valletta of Malta, 2013, 297-308

[12] 이주헌, ‘지속가능경영과 국내 외 사례’, 대한경영학회 추계학술대회, 한국, 2017, 165-175

[13] 차지이, 윤호근, 김미숙, ‘지속가능경영보고서를 이용한 ESG 요소 정성적 분석’, 한국컴퓨터종합학술대회, 한국, 2021, 1753-1755

[14] David M Blei and John D Lafferty, ‘Dynamic topic models, Proceedings of the 23rd international conference on Machine Learning, Pittsburgh of United States,2006, 113-120