Sentence BERT를 활용한 기업 SWOT 분석 자동화 연구

정상현*†, 김용신*, 정광필*, 이태희*, 권태완*

Automation of Company SWOT Analysis Using Sentence BERT

Sanghyeon Jung*†, Yongshin Kim*, Kwangpil Jeong*, Taehee Lee*, and Taewan Kwon*

요 약

클라우드 기반의 서비스형 인공지능(AI as a Service)으로 인해 고도의 인공지능 기술에 대한 접근성이 낮아짐에 따라 다양한 분야에서 지능형 애플리케이션 연구 개발이 이루어지고 있다. 본 연구는 기업 SWOT 분석을 지능적으로 자동화 할 수 있는 서비스형 인공지능인 SWOT Sentence BERT를 제안한다. SWOT Sentence BERT는 자연어 추론 형태로 가공된 SWOT 분석 데이터를 통해 학습되는 문장 임베딩 모델이며, 이를 통해 임베딩된 문장을 K-Means 알고리즘으로 클러스터링하여 기업 SWOT 분석을 자동화한다.

Abstract

As cloud-based AIaaS(AI as a Service) makes advanced AI technologies accessible, researchers in diverse disciplines have started to focus their research on developing intelligent applications. This study presented SWOT Sentence BERT as an AIaaS model that can intellectually automate company SWOT analysis. The SWOT Sentence BERT is a sentence embedding model that is learned through SWOT text data processed in the form of natural language inference task. In order to automate SWOT analysis, we applied K-Means clustering algorithm to make clusters with sentence embeddings and classified sentence embeddings based on their predicted clusters.

Key words

AI as a Service, SWOT analysis, Sentence BERT, Cloud SaaS Platform

Ⅰ. 서 론

클라우드 기반의 서비스형 인공지능(AI as a Service)은 인공지능 기술을 직접 개발하지 않고도 애플리케이션에 인공지능 기술을 도입할 수 있다는 장점으로 다양한 분야에서 지능형 애플리케이션 연구 개발에 도입되고 있다[1]. 가령, 마케팅 분야에서는 서비스형 인공지능을 기계적(Mechanical AI), 사고적(Thinking AI), 감각적(Feeling AI)으로 기능을 세분화 하여 마케팅 분석 및 애플리케이션 개발에 접목하는 연구 개발 사례가 증가하고 있다[2].

본 연구는 추후 클라우드 기반 서비스형 인공지능 플랫폼에서 상용화 될 수 있는 마케팅 인공지능 서비스 개발의 일환으로, 인공지능을 통해 SWOT 분류를 자동화하는 방식을 제안한다. 특히, 문장 쌍의 관계를 이용해서 문장의 표현을 학습하는 방식인 자연어 추론(Natural Language Inference) 기술과 비지도학습 방식인 클러스터링을 Sentence BERT모델[3]과 K-Means 알고리즘으로 각각 구현하고, 이들을 문장 분류에 적용한 결과를 지도학습 방식인 BERT[4]의 문장 분류 결과와 비교 분석한다.

* 오케스트로(주) 인공지능연구소
sh.jung@okestro.com, ys.kim2@okestro.com, kp.jeong@okestro.com, th.lee@okestro.com, tw.kwon@okestro.com
† 교신저자

※ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00147, 멀티-하이브리드 SaaS 솔루션 통합관리 플랫폼 기술 개발)

Ⅱ. 자연어 추론을 이용한 SWOT SentenceBERT의 학습

본래 자연어 추론이란, 두 문장이 주어졌을 때 두 문장 사이의 관계가 참, 거짓, 중립 중 어떤 범주인지를 예측하는 태스크로, 대표적인 국문 데이터셋으로는 KLUE NLI 데이터가 있다[5]. 본 연구에서는 이를 응용하여, 두 문장이 주어졌을 때 해당 두 문장이 강점-약점(S-W), 강점-기회(S-O), 강점-위협(S-T), 약점-기회(W-O), 약점-위협(W-T), 기회-위협(O-T)의 총 여섯 가지 관계 중 어떤 범주인지를 예측하는 태스크로 변형하였다.

본 연구에서는 이를 위해서 총 1,358개 기업의 강점, 약점, 기회, 위협 텍스트 데이터를 자체 수집하였고, 이들을 조합하여 자연어 추론 태스크 형태의 데이터셋을 구축하였다.

두 문장 간의 관계를 학습하기 위해서, 두 문장 간의 관계 학습을 통해 BERT의 문장 임베딩 성능을 우수하게 개선시킨 SentenceBERT를 활용하였다. 이와 같은 기업 SWOT 분석 자동화를 위한 SWOT Sentence BERT의 도식은 그림 1과 같다.

이처럼 두 문장 간 관계를 여섯 가지 범주로 분류하는 자연어 추론 방식으로 학습을 마친 SWOT Sentence BERT는 문장의 표현 벡터를 추출하는 문장 임베딩 벡터 추출기 역할을 한다.

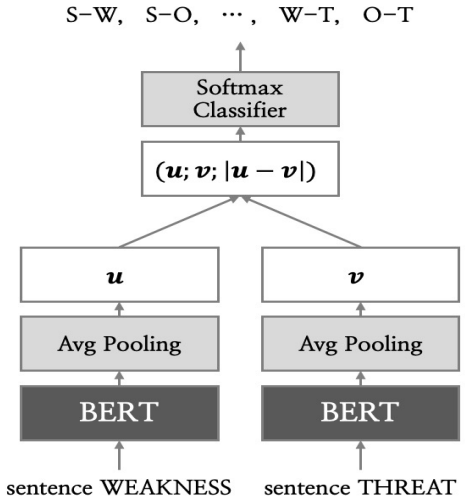

그림 1. 자연어 추론을 이용한 SWOT Sentence BERT
Fig. 1. SWOT Sentence BERT using NLI task

Ⅲ. K-Means 클러스터링을 이용한 문장 분류

본 연구에서는 SWOT SentenceBERT를 학습하기 위해 활용했던 자연어 추론 형태의 데이터를 재가공하여 단일 문장을 강점, 약점, 기회, 위협 범주로 분류하는 다중 클래스 분류 데이터셋으로 변형하였다.

이후, 정확하게 학습된 SWOT SentenceBERT는 강점, 약점, 기회, 위협 텍스트를 단일 문장으로 입력하여도 문장의 표현 벡터를 잘 포착할 수 있을 것이라는 가설에 따라, 클러스터의 개수를 네 개로 설정하였으며, 앞서 다중 클래스 분류 데이터셋으로 변형된 학습용 데이터를 활용하여 K-Means 클러스터링을 수행하였다.

이 과정에서 형성된 각 네 개의 클러스터에 소속된 학습용 데이터의 실제 정답 레이블 중 최빈값을 해당 클러스터의 레이블로 간주하였다. 이는 K-Means 클러스터링이 비지도학습 방식이라는 점을 고려하여, 테스트용 데이터의 클러스터가 예측되었을 때, 예측된 클러스터에 대해 특정 레이블을 부여하기 위함이다.

최종적으로 다중 클래스 분류 데이터셋으로 구축된 테스트용 데이터셋을 활용하여, 테스트 문장 임베딩으로 예측된 클러스터에 부여된 레이블과 해당 문장의 실제 레이블을 비교하는 방식으로 분류 성능을 측정하였다.

Ⅳ. 실험 결과

표1은 학습에 사용한 기업 데이터의 수가 일정 수준(본 실험에서는 약 70개) 이상을 사용했을 때부터 SWOT Sentence BERT의 SWOT 분류 F-1 점수가 BERT의 F-1점수보다 증가하는 결과를 보여준다.

그림 2를 통해 이러한 결과를 해석할 수 있다. 즉, 사용한 기업 데이터의 수가 일정 수준인 70개 이상이 되었을 때부터 클러스터의 적정 군집 수를 의미하는 관성(Inertia)이 4로 결정되는 것을 확인할 수 있다. 이는 SWOT Sentence BERT가 최초 여섯 개의 클래스를 분류하는 자연어 추론 방식으로 학습되었음에도 불구하고, 이를 통해 추출한 문장 임베딩의 클러스터는 정답 레이블인 Strength, Weakness, Opportunity, Threat의 특징을 반영하여 4개로 형성되기 때문이다.

V. 결 론

본 논문은 자연어 추론 방식으로 SWOT 데이터를 학습한 후 비지도학습 방식의 K-Means 클러스터링을 이용하는 SWOT Sentence BERT가 일정 수준의 데이터가 확보될 경우 지도학습 방식의 BERT보다 SWOT 분류 자동화 성능이 우수함을 보여준다. 추후 본 연구 내용을 기반으로 클라우드 플랫폼을 통해 서비스형 인공지능으로 상용화한다면, 마케팅 분석가의 수고가 들어가는 자연어 이해 기반의 SWOT 분류 작업을 자동화할 수 있는 애플리케이션으로 개발될 수 있을 것이다.

참 고 문 헌

[1] Lins, Sebastian, et al. "Artificial Intelligence as a Service." Business & Information Systems Engineering 63.4 (2021): 441-456.

[2] Huang, Ming-Hui, and Roland T. Rust. "A strategic framework for artificial intelligence in marketing." Journal of the Academy of Marketing Science 49.1 (2021): 30-50.

[3] Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[4] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[5] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).

표 1 SWOT 분류 실험 결과 (F-1 점수)

Table 1. Experimental results of the SWOT classification (F-1 score)

	Strength		Weakness		Opportunity		Threat		Weighted AVG	
# of Co.	BERT	SBERT	BERT	SBERT	BERT	SBERT	BERT	SBERT	BERT	SBERT
30	**0.8889**	0.75	0.6667	**0.7273**	0.8571	0.8571	0.7273	0.8	**0.7803**	0.7795
50	0.8421	**0.9412**	**0.9412**	0.8182	0.75	**0.8889**	**0.875**	0.6667	**0.8652**	0.8224
70	0.82	**0.93**	0.82	**0.9**	0.62	**0.83**	0.81	**0.97**	0.78	**0.93**
300	0.9524	**1**	0.907	**0.9545**	0.9048	**1**	0.88	**0.9615**	0.9093	**0.9774**
500	0.9536	**1**	0.8828	**0.9655**	0.8788	**1**	0.8421	**0.9682**	0.8897	**0.9827**
700	0.9321	**0.978**	0.8427	**0.9787**	0.8634	**0.9885**	0.8393	**0.9677**	0.8709	**0.9777**
1358	0.9261	**0.9944**	0.8815	**0.97**	0.8644	**0.9939**	0.8305	**0.971**	0.8745	**0.9813**

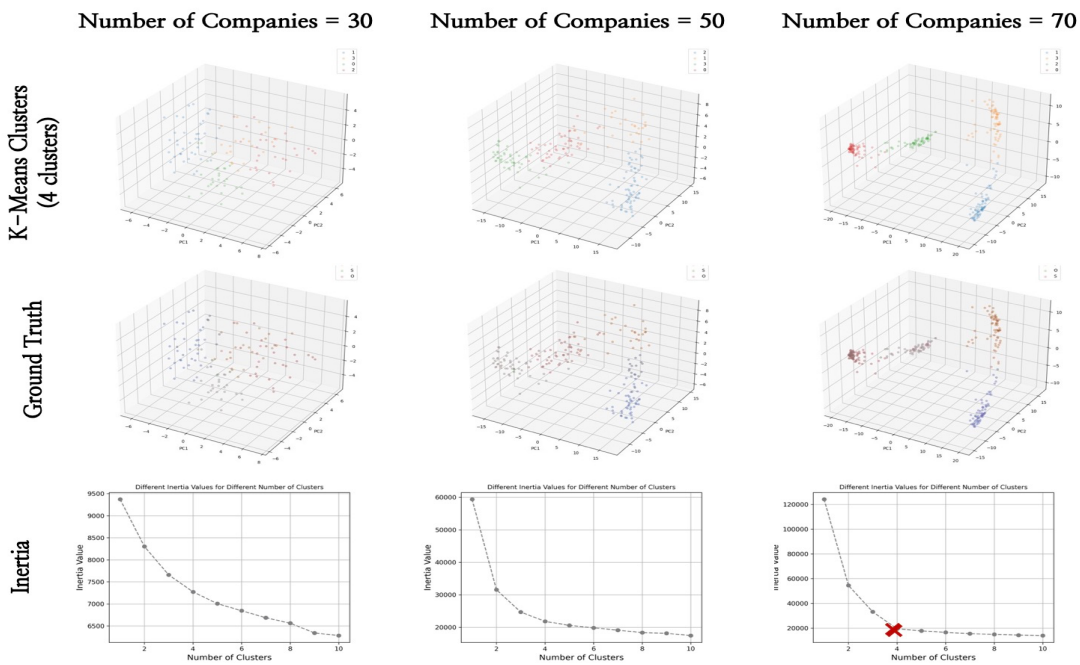

그림 2. SWOT 데이터 K-Means(n_clusters=4) 클러스터(상), 실제 정답 데이터의 분포(중), 클러스터링 관성(Inertia)(하)

Fig. 2. K-Menas clusters of SWOT data(Top) the distribution of ground truth dataset(Middle), inertia values(Bottom)