Journal of the Korean Institute of Industrial Engineers
Vol. 47, No. 4, pp. 342-350, August 2021.
ISSN 1225-0988 | EISSN 2234-6457
https://doi.org/10.7232/JKIIE.2021.47.4.342

<Original Research>

구어체 적응 사전 학습을 통한 한국어 감정 분류 성능 향상

이정훈 · 김동화 · 노영빈 · 강필성†

고려대학교 산업경영공학부

Improving Korean Emotion Classification via Colloquial-Adaptive Pretraining

Junghoon Lee · Donghwa Kim · Youngbin Ro · Pilsung Kang

School of Industrial Management Engineering, Korea University

Language models (LMs) pretrained on a large text corpus and fine-tuned on a task data have a remarkable performance for document classification task. Recently, an adaptive pretraining method that re-pretrains the pretrained LMs using an additional dataset in the same domain with the given task to make up the domain discrepancy has reported significant performance improvements. However, current adaptive pretraining methods only focus on the domain gap between pretraining data and fine-tuning data. The writing style is also different because the pretraining data, e.g., Wikipedia, is written in a literary style, but the task data, e.g., customer review, is usually written in a colloquial style. In this work, we propose a colloquial-adaptive pretraining method that re-pretrains the pretrained LM with informal sentences to generalize the LM to colloquial style. We verify the proposed method based on multi-emotion classification datasets. The experimental results show that the proposed method yields improved classification performance on both low- and high-resource data.

Keywords: Natural Language Processing, Transfer Learning, Adaptive Pretraining, Multi-Emotion Classification

1. 서 론

감성 분석(sentiment analysis)은 사람이 작성한 텍스트가 어떤 극성을 갖고 있는지, 즉 긍정인지 부정인지, 혹은 중립인지를 분류하는 것을 목적으로 하는 자연어처리의 분야이다(Liu, 2012). 감정 분석은 최근 트위터, 페이스북과 같은 SNS가 활성화됨에 따라 소비자들이 제품 후기 등을 자발적으로 포스팅하고 회사는 이를 통해 자신들의 제품에 대한 소비자들의 반응을 분석함으로써 향후 전략을 수립하는 데 이용한다(Agarwal *et al*., 2011). 최근에는 행복, 슬픔, 공포 등과 같은 감정(Emotion)에 대해 텍스트를 분석하려는 다중 감정 분류(multi-emotion classification)에 대한 연구가 활발히 이루어지고 있다(Kant *et al*., 2018).

기계학습에서 감성 분석 혹은 다중 감정 분류와 같은 문서 분류(document classification)는 합성곱 신경망(Convolutional Neural Network, CNN)과 순환신경망(Recurrent Neural Network, RNN)을 이용한 지도학습(supervised learning) 방법이 주로 사용되어 왔다(Zhang *et al*., 2015; Zhou *et al*., 2016). 최근에는 레이블이 존재하지 않는 대량의 문서를 이용해 언어 모델(language model)을 사전 학습(pretraining)한 뒤 문서 분류 데이터로 미세 조정(fine- tuning)을 수행하는 전이 학습(transfer learning) 방식이 높은 성능을 보이고 있다(Howard and Ruder, 2018; Peters *et al*., 2018). 트랜스포머(Transformer) 구조(Vaswani *et al*., 2017)를 이용한 언어 모델인 BERT(Devlin *et al*., 2018)는 대용량의 영문 위키피디아와 도서를 이용해 사전 학습을 함으로써 높은 전이 학습 성능을 기록하였다. 한글 문서로 사전 학습된 모델로는 SKT에서 구축한 KoBERT나 한국전자통신연구원에서 구축한 KorBERT 같은 모델이 존재한다.

이 논문은 2021년도 정부의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(NRF-2019R1F1A1060338)의 성과물이며, 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00471, 모델링 & 최적화 기반 오류-free 정보인프라 자율제어 기술 개발).
† 연락저자 : 강필성 교수, 02841 서울시 성북구 안암로 145 고려대학교 산업경영공학부, Tel : 02-3290-3383, Fax : 02-929-5888,
E-mail : pilsung_kang@korea.ac.kr
2021년 2월 18일 접수; 2021년 4월 23일 수정본 접수; 2021년 4월 27일 게재 확정.

사전 학습된 언어 모델은 대량의 문서로 학습되어 일반화된 언어 표상(generalized language representation)을 갖고 있지만 미세 조정에 사용되는 과업 데이터의 도메인(domain)에 대해서는 제대로 학습되지 않은 경우가 존재한다. 예를 들어 사전 학습된 언어 모델에 감성 분석 과업을 학습하기 위해 영화에 대한 긍정 및 부정 리뷰 데이터로 미세 조정을 수행할 때 사전 학습 과정에서 언어 모델이 영화에 대한 문서를 많이 학습하지 않았을 경우 해당 언어 모델은 영화 도메인에 대해 과소적합(underfitting)되었을 수 있다. 이를 해결하기 위해 미세 조정 데이터와 동일한 도메인에 속하는 문서를 이용해 사전 학습된 언어 모델에 추가로 사전 학습을 수행하는 적응 사전 학습(adaptive pretraining) 방식이 제안되어 의학, 과학, 제품 리뷰 등의 도메인을 갖는 과업에서 높은 성능을 이룩하였다(Beltagy *et al*., 2019; Gururangan *et al*., 2020; Lee *et al*., 2020). 그러나 적응 사전 학습은 주로 사전 학습과 미세 조정에 사용되는 데이터 도메인 간 차이를 보완하기 위해서 적용이 이루어졌으며 두 데이터 간의 문체의 차이에 대한 연구는 수행된 적이 없는 실정이다.

한글 문서를 사용해 사전 학습되어 공개된 BERT모델들은 주로 한글 위키피디아 문서나 뉴스 기사를 이용해 사전 학습되었다. 이러한 문서들은 다양한 도메인의 정보를 포함하고 있는 장점이 있으나 대부분 문어체로 작성되어 있다는 한계점이 존재한다. 그러나 감성 분석 혹은 감정 분류 학습에 사용되는 데이터는 주로 소비자가 작성한 후기나 비격식적인 인터넷 게시글, 혹은 댓글로써 주로 구어체로 서술되어 있다. 따라서 문어체로 작성된 문서 위주로 사전 학습된 BERT는 구어체 문서에 대해 충분한 학습이 이루어지지 않았을 것이며 전이학습의 효과가 비교적 적게 나타날 것이다. 이를 해결하기 위해서는 구어체 데이터만을 수집해 새로운 언어 모델을 학습시키는 것이 대안이 될 수 있다. 그러나 이는 큰 연산 자원이 필요하므로 상대적으로 작은 조직이나 개인 연구자에게는 실행 불가능한 방안이다. 반면 이미 배포된 언어 모델을 목적에 맞게 추가 학습시키는 것은 적은 연산 자원으로도 가능하다. 따라서 본 연구는 이에 초점을 맞추어 문어체 데이터로 사전 학습된 BERT가 구어체 데이터에 대해서도 일반화될 수 있도록 구어체 문서를 이용해 추가로 사전 학습을 수행하는 구어체 적응 사전 학습(colloquial-adaptive pretraining)을 제안하고 이를 통해 기존 언어 모델의 성능을 올릴 수 있음을 보인다.

본 연구는 선행 연구들이 데이터의 도메인에 초점을 맞춰 적응 사전 학습을 수행한 것과 달리 문체에 초점을 맞춰 적응 사전 학습을 수행한다. 따라서 도메인이 다르지만 문체가 유사한 데이터를 웹에서 수집하여 구어체 표현을 언어 모델에 학습시킨 후 그렇지 않은 상황에 비해서 얼마나 성능을 향상시킬 수 있는지 실험을 수행하였다. 이를 위해 구어체를 갖는 영화 및 스마트폰 어플리케이션 텍스트 약 13만 건을 웹 크롤링(web crawling)하여 한국어 BERT 언어모델에 대해 적응 사전 학습을 수행하였다. 그 후 BERT 언어모델을 다중 감정 분류 데이터셋에 미세 조정을 수행하는 방식으로 제안 방법을 검증하였다. 해당 다중 감정 분류 데이터셋은 수집된 SNS게시글 및 온라인 댓글들을 일곱 가지의 감정으로 분류하는 것을 목적으로 하며 특정 도메인에 속해 있지 않은 다양한 주제의 짧은 구어체 문장들로 구성되어 있다. 다중 감정 분류 실험 결과 구어체 문서로 적응 사전 학습이 수행된 한국어 BERT 언어모델이 구어체 적응 사전 학습을 수행하지 않은 경우와 문어체 문서로 적응 사전 학습을 수행한 경우보다 향상된 미세 조정 분류 성능을 보이는 것을 확인할 수 있었다.

본 논문의 구성은 다음과 같다. 먼저 제 2장에서 감성 분석과 다중 감정 분류, 그리고 적응 사전 학습에 대한 선행 연구를 소개한다. 이후 제 3장에서 언어 모델 사전 학습 방식과 본 연구에서 제안한 구어체 적응 학습을 설명한 뒤 제 4장에서 수행한 실험의 설계 방식을, 제 5장에서 실험 결과를 분석한다. 마지막으로 제 6장에서는 본 연구의 결론과 추후 연구 방향을 서술하였다.

2. 관련연구

2.1 감성 분석 및 감정 분류 데이터셋

감성 분석은 주로 긍정과 부정의 두 가지 극성에 대한 분류를 수행한다. 또한 긍정과 부정 모두에 해당하지 않는 중립까지 포함한 세 가지 범주의 분류를 수행하기도 한다. 문장에 대해 긍부정 레이블을 새로 획득하는 비용이 크기 때문에 감성분석에 사용하는 벤치마크 데이터셋은 주로 기존의 리뷰 데이터를 통해 구축되었다. 대표적으로 영화 리뷰 사이트인 IMDB의 영화 리뷰(Pang and Lee, 2005; Socher *et al*., 2013), 미국의 최대 전자상거래 사이트인 Amazon의 제품 리뷰(Ni et al., 2019), 그리고 미국의 지역 검색 사이트 Yelp의 레스토랑 리뷰가 있다. 이러한 데이터셋들은 소비자가 제약 없이 자유롭게 자신의 의견을 서술한 형태이므로 대부분 구어체를 띄고 있다는 특징이 있다.

Plutchik(1984)의 감정의 바퀴(wheel of emotions)는 사람의 감정을 기쁨(joy), 슬픔(sadness), 신뢰(trust), 혐오(disgust), 공포(fear), 분노(anger), 놀람(surprise), 기대(anticipation)의 여덟 가지로 구분하였다. 이를 근간으로 Mohammad *et al*.(2018)는 트위터의 글을 이용해 11개의 감정 범주를 갖는 다중 감정 분류 데이터셋인 SemEval Multidimension Emotion Dataset을 구축하였고 Kant *et al*.(2018)는 이와 유사하지만 특정 주제에 대해서만 존재하는 다중 감정 분류 데이터셋인 Company Tweet Dataset을 구축하였다. 다중 감정 분류는 이진분류 기반의 감성 분석에 비해 아직 많은 연구가 이루어지지 않고 있다.

2.2 감성 분석 모델

(1) 어휘 기반 감정 분석 모델

감정 분석 모델은 기계학습 기반 방법이 사용되기 이전 주로 어휘 기반(lexicon-based) 방법이 사용되었다. 어휘 기반 감정 분류는 사전에 구축한 범용 감정 사전을 이용하는 방식으로 긍정, 부정 단어의 개수를 세는 방법(Hu and Liu, 2004)과 사전을 기반으로 단어의 극성(polarity)을 판단하는 방법(Liu, 2012)이 사용되었다. 대표적인 감성 사전인 Senti-WordNet(Baccianella

et al., 2010)은 수작업으로 긍정, 중립, 부정 등의 극성을 각 단어별로 태깅하였다. Seo *et al*.(2017)은 그래프 기반 준지도학습(semi-supervised learning)을 사용해 소수의 태깅된 감성 사전만으로 대규모의 감성 사전을 구축하는 방법을 제안하였다.

(2) 기계학습 기반 감성 분석 모델

기계학습을 이용한 감성 분석 방법은 대표적으로 CNN을 이용한 방법(Kim, 2014; Zhang *et al*., 2015)과 RNN과 어텐션(attention) 메커니즘을 이용한 방법(Lin *et al*., 2017; Zhou *et al*., 2016)이 있다. 이러한 방법들은 레이블이 존재하는 문서만을 이용해 학습하는 지도학습 기반 방식이다. Peters *et al*.(2018)는 레이블이 존재하지 않는 대량의 문서를 통해 양방향 Long Short Term Memory Network(LSTM) 기반 언어 모델을 먼저 학습한 뒤 이를 임베딩(embedding)으로 사용하여 문서 분류기를 학습하는 전이 학습 방법을 제시하였다. Howard and Ruder(2018)는 대용량 문서를 통해 LSTM 기반 언어 모델을 학습한 뒤 해당 언어 모델을 문서 분류에 대해 미세 조정함으로써 높은 성능을 이룩하였다. 트랜스포머 구조를 이용한 언어 모델인 BERT(Devlin *et al*., 2018)는 대용량의 영문 위키피디아와 도서를 이용해 사전 학습을 함으로써 감성 분석을 포함한 자연어처리의 다양한 과업에 대해 높은 전이 학습 성능을 기록하였다. BERT와 유사한 구조에 더욱 많은 양의 데이터를 사용하고 학습 방식을 개선한 RoBERTa (Liu *et al*., 2019)는 트랜스포머 기반 언어 모델에 더욱 많은 데이터와 적절한 최적화 기법을 사용할 경우 성능이 개선됨을 보였다.

2.3 적응 사전 학습

적응 사전 학습(adaptive pretraining)은 대용량의 문서를 이용하여 사전 학습이 수행된 언어 모델에 레이블이 존재하는 과업 데이터(task data)로 미세 조정을 수행하기 전, 과업 데이터와 유사하거나 동일한 도메인을 갖는 데이터를 이용해 추가적인 사전 학습을 적용하는 방법이다. Sun *et al*.(2019)은 BERT 언어 모델에 대하여 감성 분석을 포함한 일곱 개의 문서 분류 데이터에 대해 적응 사전 학습을 수행함으로써 분류 정확도를 향상시켰다. 또한 적응 사전 학습을 수행함으로써 적은 수의 과업 데이터만으로도 높은 성능의 문서 분류기를 학습할 수 있음을 증명하였다. Gururangan *et al*.(2020)은 RoBERTa 언어 모델에 대해 동일한 실험을 수행하여 분류 정확도를 향상시켰으며 각 데이터를 의학, 컴퓨터공학, 뉴스, 리뷰의 네 개 도메인으로 분류하고 과업 데이터가 아니지만 동일한 도메인을 갖고 있는 데이터를 이용해 적응 사전 학습을 수행하는 것이 높은 성능을 보이는 것을 증명하였다. Lee *et al*.(2020)는 바이오, 의학 분야의 문서를 이용해 적응 사전 학습을 수행함으로써 해당 분야의 과업에서 일반적인 사전 학습을 수행한 것 보다 높은 성능을 이끌어냈다. 앞의 방법들은 모두 사전 학습된 언어 모델이 특정 도메인에 대해 과소적합되었음을 지적하며 도메인에 대해 추가적인 사전 학습을 수행하였다. 본 연구는 문서의 도메인보다 문체에 집중해 언어 모델이 사전 학습 과정에서 접한 적 없는 문체에 대한 적응 사전 학습을 수행한다.

3. 방법론

본 연구에서는 <Figure 1>에 나타난 절차와 같이 문어체로 쓰여진 텍스트 데이터를 통해 사전 학습된 BERT 언어 모델에 구어체 문장 데이터를 이용해서 적응 사전 학습을 수행한다. 이를 통해 문어체 문장 위주로 사전 학습된 BERT 언어 모델이 구어체 문장에 대해 일반화되도록 한다. 그 후 구어체 문장으로 이루어진 다중 감정 분류 데이터셋에 대해 미세조정을 수행한다. 제 3.1절에서는 BERT 언어 모델의 사전 학습과 미세조정 방식에 대해 설명하고 제 3.2절에서는 구어체를 이용한 적응 사전 학습 방법에 대해 설명할 것이다.

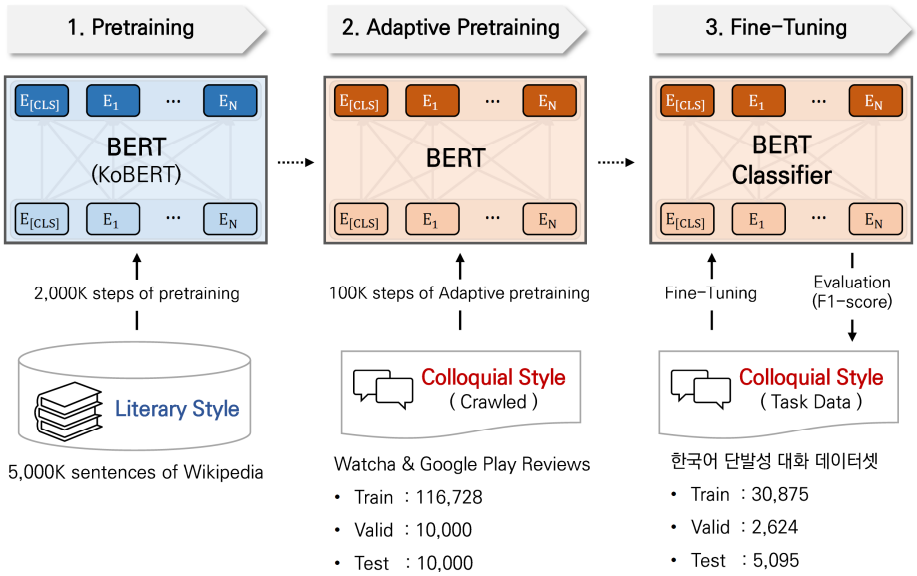

Figure 1. Research Framework

3.1 BERT 언어 모델 사전 학습

본 연구에서는 BERT의 사전 학습과 동일한 방식으로 적응 사전 학습을 수행한다. BERT 언어 모델은 기계 번역을 위해 고안된 시퀀스-투-시퀀스 모델인 트랜스포머(Vaswani *et al.*, 2017)의 인코더 구조를 차용한 언어 모델이다. 기존의 언어 모델은 양방향 순환신경망 구조를 주로 사용하였다. 그러나 순환신경망 구조는 여전히 입력 문장의 길이가 길어지면 학습이 잘 되지 않는 문제점이 존재하였다. 트랜스포머 구조는 반드시 시간의 순서에 따라 연산을 수행해야하는 순환신경망과 달리 모든 시점의 연산을 동시에 수행하기 때문에 순환신경망의 단점을 보완할 수 있었으며 GPU 병렬 연산을 더욱 효율적으로 수행할 수 있게 되어 더욱 큰 데이터셋을 이용한 학습을 가능하게 하였다. BERT는 이러한 트랜스포머의 인코더 구조를 차용하여 Masked Language Model(MLM)과 Next Sentence Prediction(NSP)라는 목적함수를 통해 언어 모델을 학습한다. <Figure 2>는 여러 층의 트랜스포머 인코더로 구성된 BERT의 구조를 나타낸 것이다.

MLM은 BERT의 입력 문장을 구성하는 토큰(token)들 중 일부를 임의로 [MASK]로 표시된 특수 토큰으로 변경한 뒤 해당 자리에 원래 있던 토큰을 예측한다. MLM은 순환신경망 기반 언어 모델에서 사용되던 단방향의 Causal Language Model (CLM)과 달리 양방향의 문맥을 모두 참고해 예측을 할 수 있다. MLM의 구체적인 식은 다음과 같다. 레이블이 존재하지 않는 문서 $X=\{x^{(1)}, \cdots, x^{(N)}\}$가 있을 때 $x=\{t_1, t_2, \cdots, t_M\}$이고 $t \in x$는 문장을 이루는 토큰이다. 이때 MLM을 학습하기 위해서는 문서 내 약 15%의 토큰을 [MASK]로 교체한다. 일부 토큰이 [MASK]로 대체된 문서를 $\hat{x}$라고 하고 원본 문서를 $\overline{x}$라고 할 때 MLM의 목적함수는 $\hat{x}$를 입력 받았을 때 $\overline{x}$을 예측하는 것이다. 목적함수를 식으로 표현하면 다음과 같다.

$$\min L_{MLM}(\theta; \hat{X}, \overline{X}) = -\sum_{\hat{x}, \overline{x} \in \hat{X}, \overline{X}} \log p_\theta(\overline{x} \mid \hat{x}) \quad (1)$$

NSP는 BERT의 입력 문장이 올바른 순서로 연결되었는지 이진 분류를 수행한다. NSP를 수행하기 위해 BERT는 두 개의 문장을 입력으로 사용한다. 이 때 일정 확률로 전혀 다른 문단에 속하는 한 문장을 선택하여 입력으로 사용한다. BERT는 위키피디아와 같은 레이블이 존재하지 않는 대용량의 문서로부터 MLM과 NSP 목적함수를 모두 사용하여 비지도학습(unsupervised learning) 방식으로 사전 학습된다. 사전 학습된 BERT 모델의 최상단에는 과업에 맞는 구조를 추가되어 미세조정을 실시함으로써 다양한 자연어처리 과업을 수행할 수 있게 된다.

3.2 구어체를 이용한 적응 사전 학습

적응 사전 학습을 수행할 때는 Liu *et al.*(2019)에서 제안한 것과 같이 NSP 목적함수를 제외하고 MLM 목적함수만 사용하였다. 적응 사전 학습을 수행하기 위해 비격식적인 구어체로 작성된 영화 리뷰 사이트 왓챠의 영화 리뷰와 구글 플레이의 스마트폰 어플리케이션 리뷰를 크롤링하였다. <Table 1>은 한국어 위키피디아, 본 연구에서 크롤링한 왓챠와 구글 플레이스토어 데이터, 그리고 다중 감정 분류 데이터의 예시를 나타내고 있다.

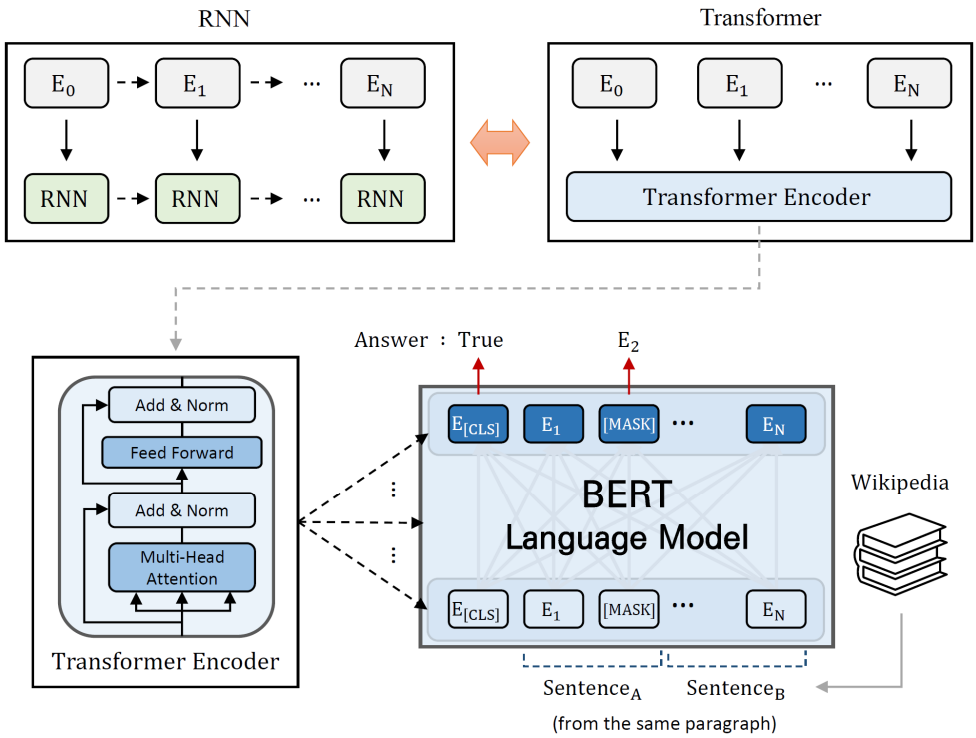

Figure 2. Illustration of Transformer Encoder and BERT Language Model

Table 1. Examples of Datasets

Source	Example
한국어 위키피디아	문어는 고정된 문법 체계를 가지고 있으므로 구어에 비해 변화가 적고 시공을 초월해 전달될 수 있다. 현대의 일상회화에서 사용되는 구어체에 비해 잘 쓰이지 않는다.
구글 플레이	와알못인데도 아예 이해가 가지않을장도로 진입장벽이 높지는 않았음 특히 차원문 앞에서의 전투신은 가히 최고라 할만함. 뜬금없는 장면들도 많았지만 그건 감독의 기량에 맡기고... 그리고 저렇게 치고박고 싸워봐야 나중에 시공의 폭풍에서 쓰까묵는거 아니겠습니까 껄껄
왓챠	정말 holy shit이란 대사가 이렇게 잘 어울리는 영화가 있을까. 정말 재미있고 소름끼친다. 연출 구성 연기 대사 엔딩 정말 다 훌륭하다. 특히 단순한 대사 한 마디에도 팀 구성원끼리의 존중 사생활 포함이 내포되어 있는 것에 감탄. 작품상 인정. 하지만 감독상은... 내 마음 속의 감독상은 매드맥스다!!!!
Multi-Emotion Classification Data	그냥 아무것도 없는 스튜디오에서 의상 가져다가 했던 우천시취소특집처럼 해줘도 정말 고마운데
	하루에도 몇번씩 추억을 찢어서 쓰레기통에 버리고이내 그걸 다시 꺼내 조각조각 맞추고 있네요..

한국어 BERT의 학습에 사용된 위키피디아의 문서들은 특정 인물 혹은 지식에 대해 설명하는 글로 이루어져 있어 대부분 문어체 형식을 띄고 있다. 반면 크롤링된 데이터는 특정 영화 혹은 어플리케이션에 대한 개인의 감상을 서술한 글로 대부분 구어체 형식을 띄고 있다. 본 연구에 사용한 다중 감정 분류 데이터셋은 게시글 및 온라인 댓글을 수집한 데이터로 비격식적인 구어체를 띄고 있다. 위키피디아의 글은 정제되면서도 많은 정보를 담고 있지만 다중 감정 분류 데이터의 예시와 같은 비격식적인 구어체는 거의 등장하지 않는다. 따라서 본 연구는 한국어 BERT를 비격식 구어체에 대해 적합한 언어 모델로 변환시키기 위해 크롤링한 구어체 데이터를 이용해 적응 사전 학습을 실시한다.

4. 실험 설계

본 논문에서 제시하는 방법론을 적용한 모델을 기존의 문서 분류 모델과 성능을 비교하는 실험을 진행하였다. 실험은 다중 감정 분류 데이터셋을 이용해 진행하였다. 추가로 가용한 데이터가 적은 실제 상황에도 제안한 방법이 유효한지 검증하기 위해 일부 데이터만 사용 가능한 상황을 가정한 실험을 진행하였다.

4.1 데이터 수집 및 전처리

본 연구에서 구어체 적응 사전 학습을 위해 사용한 데이터는 영화 리뷰 사이트 왓챠의 리뷰와 스마트폰 어플리케이션 다운로드 사이트 구글 플레이의 리뷰로 절반씩 구성된 136,728개의 데이터이다. 수집한 리뷰는 모두 5점 척도로 이루어져 있으며 1점을 부정, 3점을 중립, 5점을 긍정 레이블로 설정한 뒤 각 레이블이 동일한 비율로 존재하도록 구성하였다. 수집된 데이터는 최소 100개가 넘는 글자 수를 갖고 있다. 수집된 데이터 중 116,728개를 훈련 집합으로, 10,000개를 검증 집합으로, 10,000개를 테스트 집합으로 분리하였다.

다중 감정 분류 학습에 사용한 데이터는 한국지능정보사회진흥원이 구축한 한국어 감정 정보가 포함된 단발성 대화 데이터셋으로 행복, 중립, 놀람, 슬픔, 공포, 분노, 혐오의 일곱 가지 감정에 대한 레이블이 존재하는 38,594개의 데이터이다. 이 중 30,875개를 훈련 집합으로, 2,624개를 검증 집합으로, 5,095개를 테스트 집합으로 분리하였다. <Table 2>는 각 데이터의 세부사항을 나타낸 표이며 <Table 3>은 다중 감정 분류 데이터셋의 각 감정 범주 별 문장 예시를 나타낸 표이다. 문장 토큰화는 sentencepiece tokenizer(Sennrich *et al*., 2015)를 이용해 수행하였다.

Table 3. Multi emotion Classification Data Examples

Class	Amount	Ratio	Example
행복	6,037	15.64%	항상 밝은 에너지덕분에 힘이 납니다!!!
중립	4,830	12.51%	혈액 검사하면 금방 알 수 있다
놀람	5,898	15.28%	대박..진짜 탈퇴할 줄이야..
슬픔	5,267	13.65%	저는 아직까지 한통의 전화와 편지가 오지 않았네요ㅠ
공포	5,468	14.17%	애인으로써 정말 걱정됩니다
분노	5,665	14.68%	일본.. 절대 잊지 않겠다!!
혐오	5,429	14.07%	한 나라의 대통령으로써 부끄럽지도 않냐?

Table 2. Data Description

Data	Source	# Train	# Dev	# Test	Average lengths	# Class
Crawled Data	왓챠, 구글 플레이	116,728	10,000	10,000	167	3
Multi-Emotion Classification	한국지능정보사회진흥원	30,875	2,624	5,095	23	7

4.2 사용 모델 및 실험 환경

본 연구에서 제안한 구어체 적응 사전 학습의 효과를 검증하기 위해 한국어 위키피디아를 통해 사전학습된 KoBERT 언어 모델에 크롤링된 데이터를 사용하여 100,000 스텝(step)의 적응 사전 학습을 수행하였다. 학습에는 Adam optimizer를 사용했으며 학습률(learning rate)은 0.00005, 배치 사이즈는 64를 사용하였다. 다중 감정 분류 데이터셋에 대해 미세 조정을 수행할 때는 데이터가 적은 상황에서도 제안한 방법이 효과가 있는지를 확인하기 위해 학습 데이터 30,875개를 모두 사용할 수 있는 상황, 학습 데이터가 10,000개인 상황, 5,000개인 상황, 그리고 1,000개인 상황을 가정하여 실험을 진행하였다. 모든 데이터는 <Table 3>과 동일한 범주 비율을 갖도록 설정하였다.

제안 방법의 성능을 검증하기 위해 비교군으로 기계 학습 기반의 문서 분류 모델 네 개를 사용하였다. 첫 번째는 CNN을 이용한 문서 분류 모델(Kim, 2014)로 사전 학습된 Word2Vec 임베딩에 1차원 합성곱 필터(1D convolutional filter)를 통해 특징을 추출하고 max pooling을 수행한 뒤 선형 레이어를 통해 문서 분류를 수행한 모델이다. 두 번째는 어텐션 메커니즘을 이용한 LSTM 을 통해 문서 분류 모델(Zhou *et al*., 2016)로 앞선 모델과 마찬가지로 사전 학습된 Word2Vec 임베딩을 사용한 모델이다. 해당 모델은 양방향의 LSTM을 모두 사용함으로써 문맥을 다양하게 고려할 수 있으며 어텐션 메커니즘을 통해 문장을 구성하는 각 토큰의 중요도를 모두 고려할 수 있다. 세 번째 모델은 KoBERT에 적응 사전 학습을 수행하지 않고 바로 미세조정을 수행한 모델이다. 네 번째 모델은 KoBERT에 문어체 데이터를 이용해 적응 사전 학습을 적용한 후 미세조정을 수행한 모델(Literary-Adaptive Pretraining)이다. 문어체 적응 사전 학습에는 구어체 사전 학습에 사용한 데이터와 동일한 수의 문어체로 이루어진 뉴스 기사를 사용하였다.

4.3 성능 평가 지표

<Table 3>에서 확인할 수 있듯이 다중 감정 분류 데이터셋은 범주 별 데이터의 비중의 동일하지 않기 때문에 단순 정분류율(accuracy)보다 범주 불균형 환경에서 보다 적합한 지표인 F1-score를 평가 지표로 사용하였다. <Table 4>는 성능 평가에서 활용하는 혼동 행렬(confusion matrix)이다. 범주 i에 대하여 혼동 행렬의 행 방향은 정답 범주이고 열 방향은 모델이 분류한 범주이다. 해당 혼동 행렬을 바탕으로 식 (2)~식 (4)와 같이 정확도(accuracy), 재현율(recall), 그리고 정밀도(precision)를 구할 수 있으며 F1-score은 식 (5)와 같이 재현율과 정밀도의 조화평균으로 계산할 수 있다.

Table 4. Confusion Matrix of Category i

Category i		Classifier assignment	
		Yes	No
Human assignment	Yes	A	B
	No	C	D

$$Accuracy = \frac{A+D}{A+B+C+D} \tag{2}$$

$$Recall = \frac{A}{A+B} \tag{3}$$

$$\mathrm{Precision} = \frac{A}{A+C} \tag{4}$$

$$F1-score = \frac{2\times(Recall\times \mathrm{Precision})}{Recall+\mathrm{Precision}} \tag{5}$$

5. 실험 결과

5.1 모델 성능 평가

다중 감정 분류 테스트 데이터셋에 대한 다섯 개 모델에 대한 성능은 <Table 5>와 같다. 일곱 개의 감정 대부분에 대해서 KoBERT에 구어체 적응 사전 학습을 적용한 후 미세조정을 수행한 모델이 가장 높은 성능을 보였다. 또한 KoBERT를 사전학습할 때 사용한 문서와 동일한 문어체 형식의 문서를 사용해 적응 사전 학습을 한 경우 성능 변화가 적은 것을 확인할 수 있었다. 따라서 본 연구에서 제안한 문체 기반 적응 사전 학습이 다중 감정 분류에서 좋은 성능을 보이고 있음을 확인할 수 있었다. 기존 적응 사전 학습 방법은 미세 조정에 사용되는 데이터와 동일한 도메인을 갖고 있는 데이터를 이용해 추가적인 사전 학습을 수행함으로써 전이 학습 성능을 향상시키는 방법이었다. 본 연구는 적응 사전 학습에 사용하는 데이터가 미세 조정 데이터와 도메인이 같지 않더라도 문체가 유사하다면 효과가 있다는 점을 검증하였다.

Table 5. Model F1-Score for Each Emotion

Class	CNN	LSTM	KoBERT	Literary-Adaptive Pretraining	Colloquial-Adaptive Pretraining
행복	0.723	0.7295	0.797	0.790	**0.813**
중립	0.339	0.343	0.404	0.433	**0.451**
놀람	0.439	0.517	0.528	0.539	**0.557**
슬픔	0.569	0.483	0.599	0.607	**0.62**
공포	0.447	0.448	0.589	0.578	**0.594**
분노	0.473	0.378	0.494	**0.527**	0.519
혐오	0.285	**0.556**	0.352	0.276	0.363
평균	0.481	0.493	0.538	0.540	**0.560**

Table 6. Model F1-Score on the Low Resource Data

# Train	CNN	LSTM	KoBERT	Colloquial-Adaptive Pretraining
1,000	0.384	0.329	0.326	**0.433**
5,000	0.379	0.375	0.473	**0.537**
10,000	0.431	0.380	0.501	**0.537**
30,875	0.474	0.493	0.538	**0.560**

제안한 방법을 적용하지 않고 미세 조정을 수행한 KoBERT와 비교했을 때 제안한 방법에서 가장 높은 폭의 성능 상승을 이룬 범주는 '중립' 범주인데 이는 대부분의 모델에서 두 번째로 낮은 성능을 보이는 범주이다. 이는 제안한 방법을 적용한다면 감정 분류에서 모델이 분류를 어려워하는 중립적인 의도를 갖는 문장을 비교적 잘 분류해낼 수 있다는 것을 의미한다.

<Table 6>은 학습 데이터 수가 적은 상황일 때의 모델 성능을 나타낸다. 모든 상황에서 제안한 방법이 높은 성능을 보이는 것을 확인할 수 있었다. 제안한 방법을 적용한 후 5,000개의 학습 데이터만을 사용한 모델이 적응 사전 학습을 수행하지 않고 모든 데이터를 사용해 학습한 KoBERT와 견줄 만한 성능을 보이는 것을 통해 제안한 방법이 적은 수의 데이터인 상황에서도 큰 효과를 보이는 것을 알 수 있다.

5.2 오류 분석

<Figure 3>은 다중 감정 분류 테스트 데이터셋에 대해 제안한 방법을 적용한 KoBERT 모델의 혼동행렬이다. 학습된 모델이 '행복' 범주에 대해 가장 높은 성능을 보이며 '중립'과 '혐오' 범주에 대해 낮은 성능을 보이는 것을 확인할 수 있다. '분노' 범주에 대해서는 모델이 '혐오' 범주로 오분류를 많이

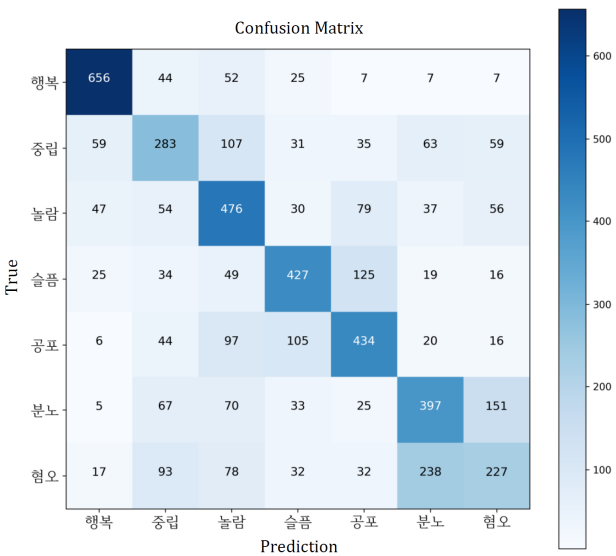

Figure 3. Confusion Matrix

Table 7. Examples of Misclassifications

Example	Prediction	True
돈있음뭐해 인성이 걸러먹었는데	혐오	분노
저런 가짜역사교육을 받고 자란 일본인이니..햐		
음주운전 단 한번이라도 차량 몰수해라	분노	혐오
대학이 기숙사 짓는데 지네가 뭔데 반대를 해!		
쇼핑하거니 끝나면 간단한 디저트하나 먹고 신나게 레이싱	행복	중립
32살인데 5년동안 방콕중이다	슬픔	놀람
백만은 상징적 숫자에 불과할 뿐이다	중립	공포
유느님 어제 지호랑 즐거운 시간보내셨겠네요	행복	슬픔

하고 반대의 경우도 동일하다. <Table 7>은 혼동행렬 상에서 모델이 예측에 어려움을 겪고 있는 범주에 대한 오분류 예시이다. 오분류된 예시를 확인해 보면 사람이 판단하기에도 정답 범주와 모델이 예측한 범주 간 구분이 쉽지 않음을 파악할 수 있다. 이는 감정이라는 범주가 한 가지로 구분되기 힘들다는 특성을 갖고 있기 때문에 일어난 상황으로 다중 감정 분류 데이터셋을 구축할 때 한 번에 여러 개의 범주를 가질 수 있도록 데이터를 구축할 필요성이 있음을 의미한다.

6. 결 론

언어 모델을 이용한 전이학습에서 미세 조정에 사용할 데이터와 동일한 도메인을 갖는 데이터를 통한 적응 사전 학습은 높은 효과를 보여 왔다. 본 논문에서는 미세 조정에 사용할 데이터와 도메인이 같지 않더라도 문체가 동일한 데이터를 이용하여 적응 사전 학습을 수행할 경우 BERT 모델의 감정 분류 성능이 향상됨을 보이고자 하였다. 이를 위해 감정 분류에 사용하는 데이터와 다른 도메인을 갖는 영화 및 스마트폰 어플리케이션 구어체 리뷰 약 13만 건을 수집하여 한국어 위키피디아 문서를 통해 사전 학습된 KoBERT 모델에 적응 사전 학습을 수행하였다. 특정한 도메인이 없는 다중 감정 분류 데이터셋에 대해 실험을 진행한 결과 기존의 CNN, LSTM, KoBERT 기반 감정 분류 모델, 그리고 사전 학습에 사용된 문서와 동일한 문체를 가진 문서를 이용해 적응 사전 학습을 수행한 KoBERT 기반 감정 분류 모델보다 높은 분류 성능을 보였다. 이를 통해 적응 사전 학습 전략은 미세 조정 데이터와 동일한 도메인에 속하는 데이터뿐만 아니라 유사한 문체를 갖는 데이터의 경우에도 효과를 보인다는 사실을 입증하였다. 또한 제안한 방법이 레이블이 존재하는 데이터가 적은 상황에서도 높은 성능을 보이는 것을 확인할 수 있었다.

본 연구에서 크롤링을 통해 수집한 데이터는 미세 조정에 사용한 데이터와 다른 출처로부터 수집되었다. 따라서 구어체를 갖고 있지만 차이가 존재한다. 그럼에도 불구하고 수집된 데이터를 이용한 적응 사전 학습이 좋은 성능을 보인 점으로 미루어 보았을 때 동일한 출처로부터 수집된 동일 문체의 데이터를 이용해 적응 사전 학습을 수행한다면 본 연구 결과보다 더욱 높은 성능 향상을 기대해볼 수 있을 것이다. 실제 산업에서 감정 분류기를 학습할 때 가장 걸림돌이 되는 것은 레이블링을 수행하는데 드는 높은 비용이다. 본 연구에서 제안한 적응 사전 학습은 레이블이 불필요한 비지도학습 방법이다. 따라서 동일한 출처로부터 수집되었던 레이블이 존재하지 않는 문서를 이용해 제안한 방법을 사용한다면 추가적인 비용 없이 감정 분류기의 성능을 더욱 향상시킬 수 있을 것이다. 또한 제안한 방법을 사용할 경우 레이블이 존재하는 데이터가 적은 상황에서도 높은 성능을 보이는 분류기를 학습할 수 있기 때문에 새롭게 감정 분류기를 구축할 때 소요되는 비용을 줄일 수 있을 것이다.

제안한 방법은 미세 조정 단계에서 사용된 과업의 종류와 상관없이 수행할 수 있는 방법론이다. 따라서 본 연구에서 진행한 감정 분류 외에도 다양한 과업의 성능을 향상시킬 수 있을 것이다. 이를 검증하기 위한 향후 연구 과제로 본 방법론이 유의미한 성능 향상을 가져올 수 있는 다른 과업에 적용해보고자 한다. 또한 본 연구에서는 구어체와 문어체만으로 문장의 종류를 구분하였다. 실제로는 구어체도 격식체와 비격식체 등으로 더욱 구체적으로 나눌 수 있다. 따라서 미세조정에 사용될 데이터의 문체를 구체적으로 판단해 문체가 더욱 비슷한 데이터를 적응 사전 학습에 사용하는 실험을 진행하여 본 연구를 고도화해보고자 한다.

참고문헌

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. J. (2011), *Sentiment Analysis of Twitter Data*, Paper Presented at the Proceedings of the Workshop on Language in Social Media (LSM 2011).

Baccianella, S., Esuli, A., and Sebastiani, F. (2010), SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining, In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, 2200-2204.

Beltagy, I., Lo, K., and Cohan, A. (2019), SciBERT : A Pretrained Language Model for Scientific Text, *arXiv preprint arXiv:1903.10676*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018), Bert : Pre-Training of Deep Bidirectional Transformers for Language Understanding, *arXiv preprint arXiv:1810.04805*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020), Don't Stop Pretraining : Adapt Language Models to Domains and Tasks, *arXiv preprint arXiv:2004.10964*.

Howard, J. and Ruder, S. (2018), Universal Language Model Fine-Tuning for Text Classification, *arXiv preprint arXiv:1801.06146*.

Hu, M. and Liu, B. (2004), *Mining and Summarizing Customer Reviews*, Paper presented at the Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Kant, N., Puri, R., Yakovenko, N., and Catanzaro, B. (2018), Practical Text Classification with Large Pre-Trained Language Models, *arXiv preprint arXiv:1812.01207*.

Kim, Y. (2014), *Convolutional Neural Networks for Sentence Classification*, Paper presented at the EMNLP.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020), BioBERT : A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining, *Bioinformatics*, **36**(4), 1234-1240.

Lin, Z., Feng, M., Santos, C. N. D., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017), A Structured Self-Attentive Sentence Embedding, *arXiv preprint arXiv:1703.03130*.

Liu, B. (2012), Sentiment Analysis and Opinion Mining, *Synthesis Lectures on Human Language Technologies*, **5**(1), 1-167.

Liu *et al*. (2019), Roberta : A Robustly Optimized Bert Pretraining Approach, *arXiv preprint arXiv:1907.11692*.

Mohammad, S., Bravo-Marquez, F., Salameh, M., and Kiritchenko, S. (2018), *Semeval-2018 Task 1 : Affect in Tweets*, Paper Presented at the Proceedings of the 12th International Workshop on Semantic Evaluation.

Ni, J., Li, J., and McAuley, J. (2019), *Justifying Recommendations Using Distantly-Labeled Reviews and Fine-Grained Aspects*, Paper Presented at the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),

Pang, B. and Lee, L. (2005), Seeing Stars : Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales, *arXiv preprint cs/0506075*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018), Deep Contextualized Word Representations, *arXiv preprint arXiv:1802.05365*.

Plutchik, R. (1984), Emotions : A General Psychoevolutionary Theory, *Approaches to Emotion*, 197-219.

Sennrich, R., Haddow, B., and Birch, A. (2015), Neural Machine Translation of Rare Words with Subword Units, *arXiv preprint arXiv:1508.07909*.

Seo, D., Mo, K., Park, J., Lee, G., and Kang, P. (2017), Word Sentiment Score Evaluation based on Graph-Based Semi-Supervised Learning and Word Embedding, *Journal of the Korean Institute of Industrial Engineers*, **43**(5), 330-340.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013), *Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank*, Paper Presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019), *How to Fine-Tune Bert for Text Classification?*, Paper Presented at the China National Conference on Chinese Computational Linguistics.

Vaswani *et al*. (2017), Attention is All You Need, *arXiv preprint arXiv:1706.03762*.

Zhang, X., Zhao, J., and LeCun, Y. (2015), Character-Level Convolutional Networks for Text Classification, *arXiv preprint arXiv:1509.01626*.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., and Xu, B. (2016), *Attention-based Bidirectional Long Short-Term Memory Networks for Relation Classification*, Paper Presented at the Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 2 : Short papers),

저자소개

이정훈 : 고려대학교 산업경영공학부에서 2019년 학사학위를 취득하였다. 현재는 고려대학교 산업경영공학부에서 석사과정으로 재학 중이다. 연구 분야는 자연어 처리이다.

김동화 : 서울과학기술대학교 글로벌융합산업공학과에서 2015년 학사학위를 취득하고, 현재는 고려대학교 산업경영공학부에서 박사과정으로 재학중이다. 연구 분야는 딥러닝을 활용한 representations learning이다.

노영빈 : 고려대학교 경영학과에서 2019년 학사학위를 취득하고, 고려대학교 산업경영공학부에서 2021년 석사학위를 취득하였다. 연구 분야는 비정형 데이터를 활용한 데이터 마이닝이다.

강필성 : 서울대학교 산업공학과에서 2003년 학사, 2010년 박사학위를 취득하였다. 이후 현대카드 과장, 서울과학기술대학교 조교수로 근무하였으며, 현재는 고려대학교 산업경영공학부 부교수로 재직 중이다. 연구 분야는 정형 및 비정형 데이터를 활용한 데이터마이닝 및 기계학습 알고리즘 개발 및 제조/IT/공공분야 응용이다.