

전이학습을 활용한 소규모 비정형 정책데이터 감성분석 모델

A sentiment analysis model for small-scale unstructured policy data using transfer learning

저자 (Authors)	안순재, 유승의, 홍순구 Soonjae Ahn, Seungui Ryu, Soongoo Hong
출처 (Source)	한국데이터정보과학회지 31(2) , 2020.3, 405-414(10 pages) Journal of the Korean Data And Information Science Society 31(2) , 2020.3, 405-414(10 pages)
발행처 (Publisher)	한국데이터정보과학회 The Korean Data and Information Science Society
URL	http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09322095
APA Style	안순재, 유승의, 홍순구 (2020). 전이학습을 활용한 소규모 비정형 정책데이터 감성분석 모델. 한국데이터정보과학회지, 31(2), 405-414
이용정보 (Accessed)	통계청 125.128.71.*** 2020/07/16 13:09 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s) and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s) for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

전이학습을 활용한 소규모 비정형 정책데이터 감성분석 모델[†]

안순재¹ · 유승의² · 홍순구³

¹²동아대학교 스마트 거버넌스 연구센터 · ³동아대학교 경영정보학과

접수 2020년 1월 21일, 수정 2020년 3월 3일, 게재확정 2020년 3월 4일

요약

최근 빅데이터 기술의 발전에도 불구하고 정책분야에서는 텍스트 등 비정형 데이터의 부족으로 감성분석 연구에 제한이 많았다. 이에 본 연구에서는 전이학습을 기반으로 소규모 비정형 정책 데이터를 활용한 감성분석 모델을 제안한다. 이를 위해 네이버 영화리뷰 20만 건의 댓글로 CNN 모델을 생성하고 지역리뷰 1만 건의 댓글을 이용하여 전이학습을 수행하였다. 분석결과 본 연구에서 제안한 전이학습 모델은 소규모 데이터만으로 생성된 모델보다 약 10%의 정확도 향상과 1epoch당 1000ms의 학습시간 단축을 보였다. 본 연구의 공헌도로 학술적으로는 한글 텍스트 감성분석에 전이학습을 처음으로 적용하여 향후 소규모 데이터의 감성분석 연구에 활용할 수 있는 이론적 기반을 제공하였다는 점이다. 실무적으로는 데이터가 부족하여 시도하기 어려웠던 정책 분야의 감성분석을 통해 정부사업의 성공여부를 판별할 수 있는 기초자료로 활용할 수 있다.

주요용어: 감성분석, 딥러닝, 소규모 데이터, 빅데이터 분석, 전이학습, 정책평가.

1. 서론

정부가 여러 정책을 원활하게 추진하기 위해서는 추진 과정에서 사업의 성과 등을 점검하고 환류하기 위한 평가 및 모니터링 방안이 필요하다. 최근 전문가에 의한 평가로 많은 시간과 비용이 수반되고 평가자의 주관에 개입될 수 있다는 문제점을 해결하기 위한 대안으로 빅데이터 분석기법이 떠오르고 있다. 특히 SNS 작성자들의 감성을 분석하는 감성분석 (Sentiment analysis) 연구가 활발하게 시행되고 있다. 감성분석은 경영학, 사회학뿐만 아니라 컴퓨터공학 등으로 그 영역이 확대 되고 있으나 (Park 등, 2007) 정책 분야에서는 감성어휘의 응용을 위한 연구가 많지 않다. Hong 등 (2019)은 부산 감천문화마을의 방문객들의 감성분석을 긍정과 부정 평가를 실시하였고 그 원인을 살펴보았다. 그들의 연구처럼 비교적 데이터가 많은 지역리뷰의 경우에는 딥러닝을 활용한 감성분석이 가능하지만 데이터가 부족할 경우에는 감성분석의 정확도가 떨어진다. 따라서 기존의 정책분야에서 딥러닝 기반의 감성분석 연구는 많지 않다. 정책 데이터를 구하기 어려운 이유는 온라인을 통한 정책토론에 시민 참여가 활발하지 않고 토론이나 댓글에 참여하는 참여자들은 대부분이 정책의 평가보다는 불만과 비속어 등을 표현하는 경우가 많기 때문이다. 이러한 문장들을 제외시키고 분석하면 연구에 적용할 수 있는 데이터의 양이 극히 감소하

[†] 이 논문 또는 저서는 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A3A2075240).

¹ 교신저자: (49236) 부산광역시 서구 구덕로 225, 동아대학교 스마트 거버넌스 연구센터, 박사.

E-mail: asj0502@donga.ac.kr

² (49236) 부산광역시 서구 구덕로 225, 동아대학교 스마트 거버넌스 연구센터, 박사.

³ (49236) 부산광역시 서구 구덕로 225, 동아대학교 경영정보학과 교수.

는 문제가 발생된다. 실제로 정책분야에서는 비정형 데이터의 계량화를 통해 시도하는 딥러닝의 경우 데이터 수집이 매우 제한적이고 데이터도 한 분야에 치중되어 있어 데이터 부족 문제가 자주 발생한다 (Bughin 등, 2018).

이러한 데이터 부족의 한계를 극복하기 위한 방법으로 전이학습 (Transfer learning)이 있다. 딥러닝 기반의 감성분석 모델에서 가장 중요한 요소는 데이터의 양이다. 기존의 연구에서는 적은 데이터를 이용하여 딥러닝 기반의 감성분석 모델을 만들기 위하여 모델의 파라미터, 입력채널 수 등을 수정해 가며 성능을 향상시키고자 하였지만 아직까지 큰 성과를 보이고 있지 않다 (Kim 등, 2016). 전통적인 딥러닝 알고리즘은 특정 과제나 문제를 해결하기 위해 학습하도록 설계되어있다. 전이학습은 사람이 여러 과제를 넘나들면서 지식을 활용하는 것이 아니라 직접적이고 한 단계 더 발전된 방식으로 데이터를 배워 나간다 (Caruana 등, 1995). 즉, 이 학습은 다른 관련된 과제에 대한 지식이나 모델을 재사용하는 방법을 사용한다. 영상처리 분야에서는 전이학습이 많이 사용되고 있었지만 (Heo와 Lim, 2019) 한글 텍스트에서는 거의 사용되고 있지 않다.

본 연구의 목적은 소규모 정책 데이터를 대상으로 전이학습을 활용한 감성분석 모델을 제안하고 모델의 타당성을 검증하는데 있다. 세부적인 연구로 첫째, 대량의 데이터를 이용하여 감성분석 모델을 생성하고 둘째, 소규모 목표 데이터를 대상으로 전이학습하여 감성분석 모델의 가중치를 조절한다. 셋째, 제안된 모델의 검증을 위하여 소량의 데이터를 사용하여 만든 감성분석 모델과 전이학습을 사용한 감성분석 모델의 정확도와 학습시간을 비교한다.

2. 선행연구

감성분석은 텍스트 마이닝의 한 분야로 자연어 처리 및 관련 컴퓨터 기술을 사용하여 텍스트의 감성을 자동으로 추출하거나 분류하는 것을 말한다 (Zhang 등, 2018). 감성분석에는 매우 다양한 응용 분야가 있다. 예를 들어 소비자 산업에서 제품리뷰 감성분석을 통해 다양한 상품에 대한 사용자 선호도를 확보하여 회사가 판매 전략을 조정하고 의사결정을 내릴 수 있다. 소셜 미디어에서 이벤트 댓글에 대한 감성분석은 사용자가 필요로 하는 정보를 빠르게 찾아내고, 의미 있는 정보를 유추해내는데 중요한 역할을 한다 (Salas-Zarate 등, 2016; Alashri 등, 2016; Mertiya와 Singh, 2016). 2000년 초 이래로 감성 분석은 자연어처리 (Natural Language Processing; NLP)에서 가장 활발한 연구 분야 중 하나가 되었다 (Hussein, 2018).

기존의 감성분석을 위한 접근방법으로는 어휘기반 (Lexicon-based approach)과 기계학습 기반 (Machine learning approach)의 방법으로 나눌 수 있다 (Chung과 Ahn, 2019). 어휘기반의 감성분석은 전통적인 감성분석 방법으로 감성사전을 구축하고 사전에 기록된 감성단어의 점수를 이용하여 감성의 극성을 판단한다 (Turney와 Littman, 2003; Hu와 Liu, 2004; Bravo-Marquez 등, 2016; Yang 등, 2013). 어휘기반 접근 방법은 적용하기 쉽지만 감성사전에 크게 의존하고 단어 사이의 위치 연결을 무시하여 매우 단순하다는 단점이 있다. 최근에는 딥러닝이 크게 발전함에 따라 딥러닝을 감성분석에 적용하는 것이 더욱 대중화되었다. Bag of Words (BoW) 과 단어 임베딩은 일반적으로 사용되는 딥러닝 모델에 사용된다. 단어 임베딩 모델의 대표적인 예는 Word2Vec이다 (Mikolov 등, 2013a; Mikolov 등, 2013b).

감성분석에 사용되는 딥러닝 방법 중 최근 가장 많이 사용되는 방법은 CNN (Convolutional Neural Network)이다. CNN은 처음에는 이미지를 처리하고 컴퓨터 비전에서 우수한 결과를 달성하는 데 사용되었다 (Lu 등, 2018). 최근 몇 년 동안 CNN은 NLP에도 똑같이 효과적이며 텍스트 분석에서 훌륭한 결과를 얻는 것으로 입증되었다 (Kim, 2014, Hong 등, 2019). Kim (2014)은 문장 수준에서 텍스트를 분류하기 위해 사전 훈련된 단어 벡터와 결합된 CNN 모형을 제안하고 모형의 우수성을 검증하였다. 하지만 한국어의 텍스트 분류 및 처리를 할 때 어절 기반의 CNN 분류기법을 적용하면 분석하기 어려운

단어 혹은 미학습 단어 (Out of vocabulary)가 등장하는 문제가 발생한다. Hong 등 (2019)은 이러한 어절 기반의 CNN 분류기법의 문제를 해결하기 위하여 형태소, 음절, 자소 단위로 분석하는 멀티채널 CNN 모델을 제안하였다.

CNN 모델은 딥러닝 방법 중 비교적 빠른 학습시간과 높은 정확도를 보이지만 매우 많은 학습데이터가 필요하다. 미디어, SNS 데이터는 많이 존재하지만 정책 분야 의 데이터는 매우 부족하여 CNN 방법을 통해 학습시키는 것이 매우 어렵다 (Bughin 등, 2018). 이러한 한계를 극복하기 위하여 최근 많은 연구에서 전이학습을 시도 하고 있다. 전이학습이란 신경망의 일부 또는 전체 신경망 가중치 파라미터를 최대우도추정 (Maximum likelihood estimation)을 통해 학습 데이터에 본격적으로 훈련시키기에 앞서, 다른 데이터 셋이나 목적함수를 사용한 후, 이를 바탕으로 본격적인 학습에서 신경망 가중치 파라미터를 더 쉽게 최적화 하는 것을 말한다. 전이학습의 가장 큰 장점은 소스 분야의 지식을 활용해서 타깃 분야의 학습을 향상시킬 수 있어 타깃 분야의 데이터가 적다면 대량의 소스 분야 데이터를 이용하여 정확도를 향상시키고 학습시간을 단축시킬 수 있다. 이미 수 년 전부터 영상처리 분야에서는 전이학습이 사용되고 있었지만 (Heo와 Lim, 2019) 최근에는 영문 텍스트 데이터를 대상으로 전이학습이 적용되고 있으나 한글 텍스트 분석에 사용된 전이학습은 없는 것으로 보인다.

Table 2.1 Transfer learning for sentiment analysis

Research	Basic model / core idea	Transfer learning methods	Datasets	Acc (%)
Sun <i>et al.</i> (2014)	GRU	Parameter -transfer	Amazon Product Review	81.08
McCann <i>et al.</i> (2017)	SDA		IMDB/SST-5	91.80/53.70
Peters <i>et al.</i> (2018)	biLM		SST-5	54.70
Dai and Le. (2015)	LSTM		IMDB	92.80
Howard <i>et al.</i> (2018)	AWD-LSTM		IMDB	95.40
Dai <i>et al.</i> (2017)	Boosting, SVM	Instance -transfer	20Newsgroups	93.15
Xu <i>et al.</i> (2011)	Transfer Self-training algorithm		NTCIR-7 MOAT Copora	73.04
Gui <i>et al.</i> (2015)	Rademacher distribution		NLP&CC 2013 cross-lingual opinion analysis dataset	80.84
Wang <i>et al.</i> (2017)	Adaptive source-domain training instance selection		Blog, Diary, Exp, Fairy	67.03
Blitzer <i>et al.</i> (2006)	Identify correspondences among features from different domains	Feature -representation	Source Domain: WSJ, Target Domain: Biomedical Text	96.10
Pan <i>et al.</i> (2010)	construct a bipartite graph to model the co-occurrence relationship		Amazon Product Review	72.86/86.75
Xia <i>et al.</i> (2013)	feature ensemble(FE), PCA-based sample selection(PCA-SS)		Amazon Product Review	72.50/84.87
Zhou G <i>et al.</i> (2015)	a joint non-negative matrix factorization		Amazon Product Review	74.83/86.33

3. 연구방법론

3.1. 학습데이터 셋 수집

본 연구에서는 전이학습을 위해 대량의 데이터 셋으로 Naver Sentiment Movie Corpus (NSMC)를 사용하였다. NSMC는 전체 20만건의 네이버 영화 리뷰 데이터 셋으로, 학습데이터 150,000건, 검증데이터 50,000건으로 구성되어 있다. 문서는 긍정 (1), 부정 (0)으로 레이블 되어 있으며 긍정 데이터 10만건, 부정데이터 10만건으로 구성되어 있다. 평점 1~4점은 부정, 9~10점은 긍정 데이터로 분류되었다. 5~8은 중립 데이터로 분류되었고 분석에서 제외되었다.

전이학습에 사용하기 위하여 세계 최대 지도 서비스인 구글맵에서 지역리뷰 데이터를 수집하였다. 한국관광청의 추천 지역명소를 파이썬 (Python)기반의 크롤러를 이용하여 총 34곳의 한국명소에 대한 지역 리뷰와 평점을 수집 하였다. 이렇게 선정된 한국명소의 지역리뷰로 총 90,542개가 수집되었고 이중 외국어로 작성된 32,178건, 한단어로 이루어진 14,755 건, 감천문화마을 지역리뷰 2,113건을 제외한 41,496건을 획득하였다. 전체 41,496건 중 긍정 5,000건, 부정 5,000건 총 10,000건의 데이터를 전이학습에 사용하기 위하여 무작위로 추출하였다. 이 중 80%인 8,000건의 데이터는 전이 학습에 사용하고 나머지 20%에 속하는 1,000건의 데이터는 모델 검증에 사용하였다. 본 연구에서 획득된 데이터는 1~3인 데이터는 부정 (0), 평점이 4~5인 데이터는 긍정 (1)으로 레이블 하였다.

3.2. 전이학습을 위한 CNN모델 생성

3.2.1. 데이터 전처리

NSMC 데이터와 지역리뷰 데이터는 모바일로 작성된 경우가 많아 오타나 신조어 등이 많이 포함되어 있다. 따라서 정확도 향상을 위해서 파이썬 기반의 맞춤법 교정기인 py-hanspell을 사용하여 수집된 데이터의 맞춤법과 띄어쓰기를 보정하였다. 형태소 벡터를 추출하기 위해 OKT (Open Korean Text) 형태소 분석기를 사용하여 단어의 토큰화가 아닌 형태소 (Morpheme) 단위로 형태소 토큰화 (Morpheme tokenization)를 수행하였다. 파이썬의 Gensim 라이브러리 사용하여 Word2Vec 모델 중 CBOW를 적용하여 CNN모델의 정확도를 비교하였다.

3.2.2. 대량 데이터 셋을 이용한 CNN 모델생성

대량 데이터 셋을 이용하여 학습할 딥러닝 모델로는 Kim (2014)의 연구에서 소개된 모델 중 다양한 데이터 셋에 대해 전체적으로 성능이 우수했던 CNN-non-static 모델을 참고하여 생성하였다. CNN 모델은 심층 신경망으로 합성곱층 (Convolution layer), 풀링층 (Pooling layer), 완전연결층 (Fully connected layer)로 구성되어 있다. 합성곱층과 풀링층에서 입력데이터의 특징 (Feature)을 추출하는 것이 CNN 모델의 가장 큰 특징이다. 사전학습 (Pretrained)을 통해 고차원의 데이터에서 특징만을 추출하여 저차원의 데이터를 얻을 수 있고, 이를 완전연결층에서 분류하여 뛰어난 성능을 보여준다. Figure 3.1은 임베딩 된 단어를 2차원 행렬로 구축하고 감성을 분류하는 합성곱 신경망의 구조를 나타내고 있다. 이 모델은 한개 합성곱층과 하나의 완전 연결층으로 구성되며, 합성곱층에서는 합성곱 연산과 맥스 풀링 연산이 차례대로 수행되어 완전 연결층의 이후에는 Softmax 연산을 통해 최종적인 감성분석을 수행한다. CNN 모델의 입력 데이터는 텍스트 데이터에서 워드임베딩을 통해 얻은 문장단위의 2차원 행렬이다. 워드임베딩 과정을 거쳐 2차원 행렬로 구성된 입력데이터가 준비되면 필터(Filter)를 통해 읽어 내려가며 입력 데이터와 필터와의 합성곱 연산으로 특징 지도 (Feature map)를 생성한다. 필터 크기는 합성곱층에서 연산을 처리할 때, 한 번에 처리할 수 있는 근접 데이터 정보의 크기라고 볼 수 있다. 사전에 정의된 입력 데이터의 크기가 결정되면, 합성곱층 필터의 최적의 크기는 실험적으로 결정하는데

본 연구에서는 Table 3.1에서 보듯이 NSMC 데이터로 필터크기를 (3,3,3)에서 (5,5,5)까지 비교하여 분류정확도가 높은 (4,3,4)를 최종으로 결정하였다 (Zeiler와 Fergus, 2014).

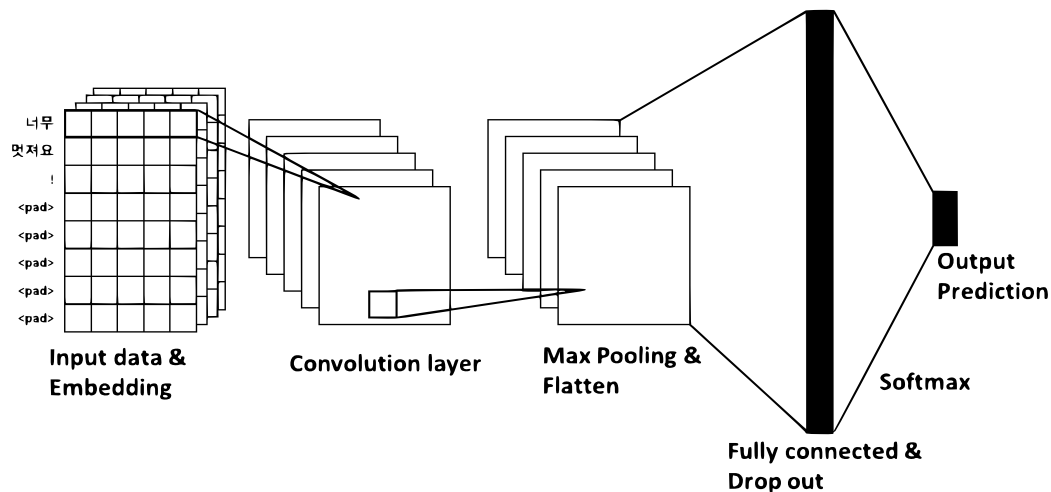


Figure 3.1 CNN model structure

Table 3.1 CNN model details

Category	Item	Value
Model hyperparameters	Embedding dimension	150
	Filter sizes	(4, 3, 4)
	Number of filters for each filter size	50
	Dropout probabilities	(0.5, 0.8)
	Hidden dimension	50
Training parameters	Batch size	64
	Number of epochs	100
Word2Vec parameters	MinimumWord count (min.count)	(1, 3, 5)
	Context (window)	5
	Vector dimension (size)	150

3.3. 전이학습

본 연구에서 제안하는 전이학습은 Figure 3.2에서 보듯이 네이버에서 제공하는 대량의 영화리뷰데이터 NSMC 20만건을 사용하여 딥러닝 모델을 생성한 후 상대적으로 적은 수의 지역 리뷰 데이터를 전이 학습시켜 모델의 가중치를 변경하는 모델이다.

3.2절에서 사전 훈련된 CNN 모델을 로드하고 가중치를 초기화 하였다. 그리고 지역리뷰 데이터 1만건을 이용하여 가중치가 초기화된 CNN 모델을 학습하고 정확도와 학습시간을 확인하였다.

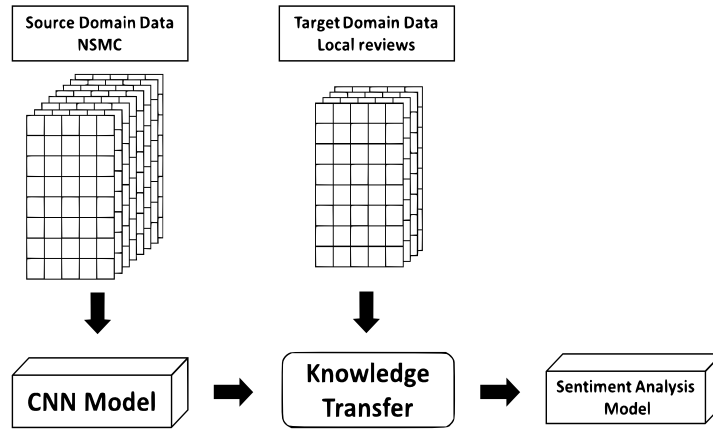


Figure 3.2 Transfer learning sentiment analysis model

4. 연구결과

실험은 window10 운영체제에서 Tensorflow 2.0 을 이용하여 진행되었으며, 딥러닝 모델의 효율적인 학습을 위해 RTX2080 GPU를 이용하였다.

4.1. 전이학습 여부에 따른 정확도 비교

수집된 데이터를 바탕으로 학습결과는 Table 4.1과 같다. 지역리뷰 4만 건의 CNN 학습결과와 1만 건의 학습결과를 비교하면 데이터양이 줄어들에 따라 73.03%에서 66.71%로 정확도가 하락하는 것을 확인하였다. 또한 대량의 데이터인 NSMC 데이터로 CNN 모델을 생성하였을 경우 정확도가 84.04%로 가장 높았다. NSMC 데이터로 학습모델을 만들고 지역리뷰 1만 건을 전이학습 시킨 경우 Table 4.1.1에서 보듯이 76.69%로 소량의 데이터로만 학습한 경우보다 정확도가 높았으며 4만 건의 지역리뷰데이터로 학습한 경우보다 정확도가 더 높았다.

Table 4.1 Training accuracy

Model	Accuracy
Local reviews CNN model (40,000 cases)	73.03 %
Local reviews CNN model (10,000 cases)	66.71 %
NSMC CNN model (200,000 case)	84.04 %
NSMC CNN model → Local reviews transfer learning	76.69 %

4.2. 전이학습 여부에 따른 학습시간 비교

정확도뿐만 아니라 학습시간에서도 차이가 발생하는지를 관찰하였다. Table 4.2에서 보듯이 지역리뷰 1만 건으로 CNN 학습할 때보다 NSMC 데이터로 학습모델을 만들어 놓고 지역리뷰 1만 건으로 전이학습을 할 때 1 epoch당 2968ms에서 1822ms로 학습시간이 단축되었다 .

Table 4.2 Training time

Model	Training time
Local reviews model	2968 ms / 1 epoch
NSMC CNN model → Local reviews transfer learning	1822 ms / 1 epoch

5. 결론

본 연구에서는 전이학습을 통하여 소규모 데이터의 감성평가 결과를 향상시키고자 하였다. 기존의 딥러닝 기반의 감성평가 방법으로는 데이터 양이 적으면 학습의 정확도가 떨어진다는 한계가 있었다. 이를 위해 본 연구에서는 네이버 영화리뷰라는 대량의 데이터를 통하여 딥러닝 모델을 만들고 이 모델을 재사용하는 방법을 택하였다. 이러한 방법을 통하여 목표로 하고 있는 지역리뷰 감성분석 모델의 학습을 향상시켰고 높아진 정확도를 확인하였다. 또한 모델의 가중치만을 변경시키는 전략을 택하여 모델을 생성하는 단계를 건너뛰어 학습시간을 단축하였다. 분석결과 대량의 데이터로 학습모델을 만들고 소량의 데이터로 전이학습을 사용하였을 때 소량의 데이터만으로 학습하였을 때보다 약 10%의 정확도 향상과 1epoch 당 약 1000ms의 학습시간 단축을 확인할 수 있었다. 이는 전이학습을 통하여 데이터 양이 적은 분야에서도 딥러닝 감성평가가 가능하다는 것을 의미한다. 또한, 기존 선행연구에서는 영어 텍스트 데이터를 이용하여 전이학습 감성분석을 시도하였지만 본 연구에서는 한글 텍스트에서도 전이학습 감성분석이 가능하다는 것을 증명하였다.

본 연구의 공헌도로 학술적으로는 한글 텍스트 감성분석에 전이학습을 처음으로 적용하여 향후 소규모 데이터의 감성분석 연구에 활용할 수 있는 이론적 기반을 제공하였다는 점이다. 실무적으로는 데이터가 부족하여 시도하기 어려웠던 정책 분야의 감성분석을 통해 정부사업의 성공여부를 판별할 수 있는 기초자료로 활용할 수 있다.

본 연구가 가지는 한계점과 향후연구 과제는 다음과 같다.

첫째, 전이학습에 대한 수학 및 이론적인 배경이 아직 정립되지 않아 전이학습을 구성할 때 상당히 경험적으로 접근할 수밖에 없다. 경험적인 접근을 통해서 더 좋은 결과를 얻을 수 있지만 정확한 이론적인 내용이 정립되어 있다면 시행착오 없이 좋은 성능을 기대할 수 있을 것이다.

둘째, 본 연구에서 얻은 결과는 네이버 영화리뷰라는 매우 많은 데이터 셋을 활용하여 평범한 CNN 모델의 성능을 향상시켜 전이학습의 가능성을 확인하였지만 더 복잡한 모델을 만들어 성능의 변화를 관찰할 필요가 있다.

셋째, 본 연구에서는 지역리뷰데이터를 통하여 소량의 데이터로 딥러닝 감성분석 모델의 성능향상을 확인하였지만 실제 정책데이터 뿐만 아니라 데이터양이 작아 딥러닝을 통한 분류작업이 어려웠던 분야에 적용한다면 도움이 될 것으로 사료된다.

References

- Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L. and Desouza, K. C. (2016). An analysis of sentiments on Facebook during the 2016 U.S. presidential election. *Proceeding of IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 795-802.
- Blitzer, J., McDonald, R. and Pereira, F. (2016). Domain adaptation with structural correspondence learning. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 120-128.
- Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2016). Building a Twitter opinion lexicon from automatically-annotated tweets. *Knowledge-Based Systems*, **108**, 65-78.

- Bughin, J., Seong, J., Manyka, J., Chui, M. and Joshi, R. (2018). *Notes from the AI frontier: Modeling the impact of AI on the world economy*, McKinsey & Company, USA, Discussion paper
- Caruana, R., Silver, D. L., Baxter, J., Mitchell, T. M., Pratt, L. Y. and Thrun, S. (1995). *Learning to learn: Knowledge consolidation and transfer in inductive systems*, NIPS*95 Post-Conference Workshop.
- Choi, S. and Yeo, J. (2010). A comparative analysis on policy evaluation methods: Focused on fair labeling & advertising act. *Survey Research*, **11**, 57-79.
- Chung, P. and Ahn, H. (2019). A sentiment analysis model for customer reviews considering sentence location. *The Journal of Internet Electronic Commerce Research*, **19**, 167-186.
- Dai, W., Yang, Q., Xue, G.-R. and Yu, Y. (2017). Boosting for transfer learning. *Proceedings of the 24th International Conference on Machine Learning*, 193-200.
- Gui, L., Lu, Q., Xu, R., Wei, Q. and Cao, Y. (2015). Improving transfer learning in cross lingual opinion analysis through negative transfer detection. *Proceedings of International Conference on Knowledge Science, Engineering and Management*, 394-406.
- Heo, K. and Lim, D. (2019). Noise reduction using patch-based CNN in images. *Journal of the Korean Data & Information Science Society*, **30**, 349-363.
- Hu M. and Liu, B. (2004). Mining and summarizing customer reviews. *Proceeding of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data*, 168-177.
- Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. *Journal of King Saud University - Engineering Sciences*, **30**, 330-338. 2018.
- Hong, S., Ryu, S. and Ahn, S. (2019). Sentimental analysis on urban regeneration policy: Focused on reviews of Gamcheon culture village. *Journal of the Korean Data & Information Science Society*, **30**, 1233-1244.
- Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification. *arXiv:1801.06146*.
- Kim, Y. (2014). Convolutional neural network for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746-1751.
- Kim, H., Yang, S., Kang, Y., Park, Y. and Kim, J. (2016). Machine learning approach to blood stasis pattern identification based on self-reported symptoms. *Korean Journal of Acupuncture*, **33**, 102-113.
- Lu, H., Li, Y., Uemura, T., Kim, H. and Serikawa, S. (2018). Low illumination underwater light field images reconstruction using deep convolutional neural networks. *Future Generation Computer Systems*, **82**, 142-148.
- McCann, B., Bradbury, J., Xiong, C. and Socher, R. (2017). Learned in translation: Contextualized word vectors. *arXiv:1708.00107*.
- Mertiya M. and Singh, A. (2016). Combining naive bayes and adjective analysis for sentiment detection on Twitter. *Proceeding of International Conference on Inventive Computation Technologies (ICICT)*, 1-6.
- Mikolov, T., Yih, W. T. and Zweig, G. (2013a). Linguistic regularities in continuous space word representations. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746-751.
- Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013b). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Pan, S. J., Ni, X., Sun, J. T., Yang, Q. and Chen, Z. (2010). Cross-domain sentiment classification via spectral feature alignment. *Proceedings of the 19th International Conference on World Wide Web*, 751-760.
- Park, D., Lee, J., and Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, **11**, 125-148.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv:1802.05365*.
- Salas-Zarate, M. P., Medina-Moreira, J., Alvarez-Sagubay, P. J., Lagos-Ortiz, K., Paredes-Valverde, M. A. and Valencia-Garcia, R. (2016). Sentiment analysis and trend detection in Twitter. *Proceeding of CITI 2016*, 63-76.
- Sun, M., Tan, Q., Ding, R. and Liu, H. (2014). Cross-domain sentiment classification using deep learning approach. *Proceeding of IEEE 3rd International Conference on Cloud Computing and Intelligence Systems*, 60-64.
- Turney P. D. and Littman, M. L. (2003). Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, **21**, 315-346.
- Wang, W., Chen, L., Chen, K., Thirunarayan, K. and Sheth, A. P. (2017). Adaptive training instance selection for cross-domain emotion identification. *Proceedings of the International Conference on Web*

- Intelligence*, 525-532.
- Xia, R., Zong, C., Hu, X. and Cambria, E. (2013). Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *IEEE Intelligent Systems*, **28**, 10-18.
- Xu, R., Xu, J. and Wang, X. (2011). Instance level transfer learning for cross lingual opinion analysis. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 182-188.
- Yang, A. M., Lin, J. H., Zhou, Y. M. and Chen, J. (2013). Research on building a Chinese sentiment lexicon based on SO-PMI. *Applied Mechanics and Materials*, **263**, 1688-1693.
- Zeiler, M. D. and Fergus. R. (2013). Visualizing and understanding convolutional networks. *arXiv: 1311.2901*.
- Zhou, G. Zhou, Y., Guo, X., Tu, X. and He, T. (2015). Cross-domain sentiment classification via topical correspondence transfer. *Neurocomputing*, **159**, 298-305.

A sentiment analysis model for small-scale unstructured policy data using transfer learning[†]

Soonjae Ahn¹ · Seungui Ryu² · Soongoo Hong³

¹²Smart Governance Research Center, Dong-A University

³Department of Management Information Systems, Dong-A University

Received 21 January 2020, revised 3 March 2020, accepted 4 March 2020

Abstract

Despite the recent development in big data technologies, the research on sentiment analysis is still facing many limitations due to the lack of unstructured data, including texts, in the policy field. Thus, this study proposes a sentiment analysis model for small-scale unstructured policy data based on transfer learning. As a result, the proposed transfer learning model achieved about 10% better accuracy and a shorter training time of 1000 ms per epoch than the model generated only by using small-scale data. As an academic contribution, this study, which is the first application of transfer learning to Korean text sentiment analysis, provides a theoretical basis for future research on sentiment analysis using small-scale data. For practicality, this study can serve as basic data in determining the success or failure of government projects through sentiment analysis in the policy field, which was difficult to determine previously given the lack of data. the detailed feature of sea/land breeze at each site is closely associated with the local shape of coastline.

Keywords: Big data analysis, Deep learning, Policy evaluation, Sentiment analysis, Small-scale data, Transfer learning.

[†] This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2018S1A3A2075240).

¹ Corresponding author: Ph.D., Smart Governance Research Center, Dong-A University, Busan 49236, Korea. E-mail: asj0502@donga.ac.kr

² Ph.D., Smart Governance Research Center, Dong-A University, Busan 49236, Korea.

³ Department of Management Information Systems, Dong-A University, Busan 49236, Korea.