JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 7, pp. 19-27, July 2023
https://doi.org/10.9708/jksci.2023.28.07.019

Improving Accuracy of Noise Review Filtering for Places with Insufficient Training Data

Hyeon Gyu Kim*

*Associate Professor, Div. of Computer Science and Engineering, Sahmyook University, Seoul, Korea

[Abstract]

In the process of collecting social reviews, a number of noise reviews irrelevant to a given search keyword can be included in the search results. To filter out such reviews, machine learning can be used. However, if the number of reviews is insufficient for a target place to be analyzed, filtering accuracy can be degraded due to the lack of training data. To resolve this issue, we propose a supervised learning method to improve accuracy of the noise review filtering for the places with insufficient reviews. In the proposed method, training is not performed by an individual place, but by a group including several places with similar characteristics. The classifier obtained through the training can be used for the noise review filtering of an arbitrary place belonging to the group, so the problem of insufficient training data can be resolved. To verify the proposed method, a noise review filtering model was implemented using LSTM and BERT, and filtering accuracy was checked through experiments using real data collected online. The experimental results show that the accuracy of the proposed method was 92.4% on the average, and it provided 87.5% accuracy when targeting places with less than 100 reviews.

▸ **Key words**: Social review, Noise review filtering, Group learning, LSTM, BERT

[요 약]

소셜 리뷰를 수집하는 과정에서 주어진 검색어와 상관없는 노이즈 리뷰가 검색 결과에 다수 포함될 수 있으며, 이들을 필터링하기 위해 기계 학습이 이용될 수 있다. 그러나 분석하고자 하는 대상의 리뷰 수가 부족한 경우, 학습 데이터 부족으로 인한 정확도 저하 문제가 발생할 수 있다. 본 논문에서는 리뷰 수가 부족한 플레이스를 대상으로 노이즈 리뷰 필터링의 정확도를 높이기 위한 지도 학습 방법을 소개한다. 제안 방법에서는 개별 플레이스 단위로 학습을 수행하지 않고, 특성이 유사한 여러 플레이스를 그룹으로 묶어 학습을 수행한다. 학습을 통해 얻은 분류기는 그룹에 속한 임의의 플레이스에 공통으로 적용함으로써 학습 데이터 부족 문제를 해결하고자 하였다. 제안 방법의 검증을 위해, LSTM과 BERT를 이용하여 노이즈 리뷰 필터링 모델을 구현하고, 온라인에서 수집된 실제 데이터를 활용한 실험을 통해 필터링 정확도를 체크하였다. 실험 결과, 제안 방법의 정확도는 평균 92.4% 수준이었으며, 리뷰 수가 100개 미만인 플레이스를 대상으로 할 경우 87.5%의 정확도를 제공하였다.

▸ **주제어**: 소셜 리뷰, 노이즈 리뷰 필터링, 그룹 학습, LSTM, BERT

• First Author: Hyeon Gyu Kim, Corresponding Author: Hyeon Gyu Kim
*Hyeon Gyu Kim (hgkim@syu.ac.kr), Div. of Computer Science and Engineering, Sahmyook University
• Received: 2023. 06. 09, Revised: 2023. 06. 27, Accepted: 2023. 06. 29.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

SNS 피드 및 블로그 리뷰 등을 포함한 소셜 빅데이터는 고객 관점의 의견이나 불만 사항을 추출하기 위해 대중적으로 이용되고 있다[1, 2]. 소셜 빅데이터는 텍스트 형태로 구성되어 신용카드나 휴대폰 이용 내역 등 수치로 이루어진 데이터에 비해 의견이나 불만 사항의 직접적인 원인을 바로 알아낼 수 있다는 점에서 선호된다. 또한 구글이나 네이버 등의 온라인 포털 업체들이 제공하는 오픈 API[3, 4]를 통해 무료로 획득 가능한 장점도 있다.

오픈 API를 이용하여 수집된 리뷰에는 주어진 검색어와 연관성이 떨어지는 소위 “노이즈” 리뷰가 다수 포함될 수 있다. 이들 중 대다수는 검색어로 주어진 플레이스가 아닌 다른 플레이스를 설명하기 위해 검색어가 차용되고 있는 경우에 해당한다. 예를 들어, 검색어로 울산 지역의 인기 식당인 “함양집”이 주어졌을 때, 아래와 같은 리뷰가 검색 결과에 포함될 수 있다.

- 함양집 건너편에 위치한 깔끔하고 넓은 매장의 빅스트로 뷔페 …
- 함양집 건물 2층에 위치한 분위기 카페 커피반하다 감성 낭낭한 …

문제는 이들 리뷰가 검색 결과에서 차지하는 비율이 높다는 점이다. 식당의 경우 약 55%의 리뷰가 주어진 검색어와 연관성이 떨어지는 것으로 조사되었다(3장 참조). 따라서 소셜 빅데이터 분석 응용에서 정확한 결과를 얻기 위해서는 수집된 데이터로부터 노이즈 리뷰를 필터링하기 위한 작업이 반드시 선행되어야 한다.

이전 연구[5]에서 논의한 바와 같이, 노이즈 리뷰는 지도 학습을 이용하여 필터링될 수 있다. 단, 지도 학습을 통해 만족할만한 정확도를 얻기 위해서는 학습 데이터가 충분해야 한다. 학습 레코드 수가 100개 이하인 경우에는 의미 있는 정확도를 얻기 어려운 것으로 알려져 있다. 반면, 소상공인이 운영하는 사업장의 경우, 대다수가 대중들에게 알려져 있지 않아 학습에 필요한 리뷰 수가 부족할 수 있다. 예를 들어 울산광역시의 경우 16,900여개의 식당이 등록되어 있으며, 이 중 64%의 리뷰 수가 100개 이하인 것으로 조사되었다(3장 참조). 이는 식당뿐만 아니라 카페, 미용실 등 600만 소상공인이 운영하는 대다수의 사업장이 지닌 공통적인 문제로 볼 수 있다.

본 논문에서는 리뷰 수가 부족한 플레이스를 대상으로 노이즈 리뷰 필터링의 정확도를 높이기 위한 지도 학습 방법을 소개한다. 제안 방법의 기본 아이디어는 동일한 지역 및 사업 카테고리(식당, 카페, 미용실 등)에 속한 플레이스들을 그룹으로 묶어 학습을 수행하는 것이다. 그룹에 속한 플레이스들은 유사한 속성을 가지므로, 이들의 일부를 샘플링하여 학습시킨 분류기(Classifier)는 그룹 내 다른 플레이스의 리뷰 필터링에 공통으로 적용될 수 있다. 따라서 리뷰 수가 부족한 플레이스의 학습 문제도 제안 방법을 통해 해결될 수 있다.

본 논문에서는 그룹 학습과 관련한 초기 연구로써, 그룹 학습의 효율성 및 정확도를 최대화할 수 있는 구체적인 방법이나 이론적인 체계를 찾기 보다는, 만족할만한 성능을 제공하는 그룹 학습 방법이 실제 존재하는지 실험을 통해 확인하고자 하였다. 실험을 위해, 울산광역시에 등록된 16,900여개의 식당을 대상으로 네이버에서 제공하는 검색 API[3]를 이용하여 약 154만개의 리뷰를 수집하였으며, 이 중 114개의 플레이스의 25,000여개의 리뷰를 샘플링하여 지도 학습을 수행하였다. 학습 알고리즘으로는 텍스트 처리에 높은 정확도를 제공하는 것으로 알려진 LSTM(Long Short-Term Memory)[6]과 BERT(Bidirectional Encoder Representation for Transformers)[7]가 이용되었다.

본 논문의 구성은 다음과 같다. 2장에서는 리뷰 필터링과 관련한 기존 연구들에 대해 소개한다. 3장에서는 실험을 수행하기 위해 수집된 데이터의 특성을 소개하고, 노이즈 리뷰 필터링을 수행하기 위한 LSTM과 BERT의 학습 모델에 대해 설명한다. 4장에서는 기존 학습 방법과 제안한 그룹 학습 방법의 정확도를 실험하고 결과를 비교한다. 리뷰 수가 부족한 플레이스를 대상으로 제안 방법의 효용성을 검증하기 위한 실험 결과도 함께 논의한다. 5장에서는 결론 및 추후 연구 방향 제시로 마무리한다.

II. Related Work

[8]에서는 스팸 리뷰를 광고, 브랜드 분석 리뷰, 특정 제품을 홍보하거나 폄하하기 위한 리뷰 등으로 분류했으며, 특히 세 번째 형태의 리뷰의 경우 사용자들을 기만하기 위해 리뷰가 진짜처럼 보이도록 신중하게 작성되었기 때문에 전통적인 기계학습 솔루션만으로는 한계가 있음을 지적하였다. 이로부터 [9]에서는 리뷰에 작성된 콘텐츠 이외에 해당 리뷰를 작성한 사용자의 동적 행위를 분석하여 필터링의 정확도를 높이고자 하였다. [10]에서는 판매량, 가격, 별점 등의 상품 관련 정보를 활용하여 필터링에 활용하였으며, [11]에서는 스팸을 보내는 사용자 그룹의 유사성을 필터링에 반영하여 정확도를 높이고자 하였다.

최근 들어 딥러닝 기술의 발전으로 인해, 별도의 행위 분석이나 부가 정보 없이 리뷰에 포함된 텍스트 분석만으로 고도의 정확도를 얻을 수 있게 되었다. [12]는 딥러닝의 가장 기본적인 알고리즘인 심층 신경망(DNN, Deep Neural Network)을 이용하여 89%의 필터링 정확도를 구현하였다. [13]에서는 CNN(Convolutional Neural Network)[14]과 LSTM 각각을 활용한 스팸 리뷰 필터링 정확도를 실험하였으며, LSTM이 CNN에 비해 1~2% 정도 우수한 정확도를 제공하는 것으로 확인되었다. [15, 16]은 트랜스포머를 기반으로 광범위한 데이터를 사전 훈련하여 만든 BERT를 이용하여 스팸 리뷰 필터링을 구현하고자 하였다. 위 방법들은 텍스트 형태의 리뷰를 수치 형태의 벡터로 표현하기 위해 Word2Vec[17]와 같은 단어 임베딩 기법을 이용한다는 공통점이 있다.

최근 연구에서는 하나 이상의 딥러닝 알고리즘을 연결하여 필터링 정확도를 높이기 위한 시도도 함께 이루어졌다. [18]은 학습 성능 향상을 위해 전체 리뷰 중 일부 특징(Aspect)만 추출하여 학습 데이터로 이용하였으며, CNN과 LSTM을 연결한 형태로 학습 모델을 구현하였다. [19]는 셀프-어텐션(Self-attention)과 CNN, LSTM을 함께 연결하여 필터링을 구현하였으며, 평균 93.6%의 높은 정확도를 제공하였다.

단, 위에서 언급된 방법들이 노이즈 리뷰 필터링에도 효과적으로 적용될 수 있을지는 확인된 바가 없다. 또한 리뷰 수가 부족한 경우 필터링 정확도를 높이기 위한 방법과 관련한 기존 연구도 찾기 어려운 실정이다.

III. Experimental Settings

1. Data Collection and Sampling

본 논문에서는 울산광역시에 등록된 플레이스를 대상으로 실험을 진행하였다. 울산광역시는 남구, 동구, 북구, 울주, 중구 등의 5개의 지역구로 구성되어 있으며, 대한민국 7개의 광역시 중 규모가 가장 작아 실험을 위해 수집해야 할 데이터의 크기가 비교적 작다는 점에서 실험 대상으로 선정되었다. 플레이스 정보는 소상공인진흥공단[20]에서 제공하는 공공 데이터를 통해 얻을 수 있었으며, 여러 카테고리의 플레이스 중 식당을 실험 대상으로 선정하였다.

수집된 데이터에는 16,900여개의 식당 정보가 포함되어 있었으며, 최근 코로나 사태 등으로 인한 폐업이나 이전과 관련된 정보가 데이터에 반영되지 않은 경우가 많았다. 따라서 공공데이터에 포함된 각각의 식당이 실제 존재하는지 확인하기 위한 보정 작업이 수행되었다. 이를 위해 네이버의 주소 검색 API[3]가 이용되었으며, 각각의 식당에 대해 API를 통해 획득한 주소와 공공데이터에 명세된 주소가 일치하는지 확인하였다.

주소 검증 결과, 전체 플레이스 중 7,445개의 플레이스가 유효한 것으로 확인되었으며, 이들을 대상으로 네이버의 블로그 검색 API를 활용하여 리뷰 수집 작업을 진행하였다. 리뷰 수집은 2023년 1월에 수행되었으며, 유효 플레이스를 대상으로 수집된 사용자 리뷰 수는 약 154만개였다.

먼저 유효 플레이스를 대상으로 리뷰 수에 따른 플레이스 분포를 확인하고자 하였다. 네이버 API를 이용할 경우 플레이스 별로 매월 수집될 수 있는 리뷰 수는 최대 1,000개이므로, 이로부터 리뷰 수 구간을 1000, 500, 250, 100, 50을 기준으로 다섯 개로 나누고, 각 플레이스를 리뷰 수에 따라 해당 구간의 그룹으로 분류하였다. 플레이스 집합 P에서 i번째 플레이스를 p_i라 하고, p_i의 리뷰 수를 n_i라고 하자. p_i는 그룹 G_b에 다음과 같이 할당될 수 있다.

$$p_i \in G_b \text{ if } \lfloor b/2 \rfloor_{50} < n_i \leq b \quad (1)$$
$$\text{where } b \in \{1000, 500, 250, 100, 50\}$$
$$\lfloor x \rfloor_{50} = x - x \bmod 50$$

식 (1)에서 $\lfloor x \rfloor_{50}$은 50의 배수 중 x를 넘지 않는 가장 큰 수를 의미하며, $x - x \bmod 50$으로 계산될 수 있다. 예를 들어 b가 250일 경우, $\lfloor 250/2 \rfloor_{50}$은 100이 된다. 따라서 G_{250}은 리뷰 수가 250 이하이며 100보다 큰 플레이스들의 집합이 된다. 아래 표는 울산광역시의 각 구별로 G_b에 포함된 식당 수를 보여준다.

Table 1. Number of the restaurants included in group G_b for each district of the Ulsan metropolitan city

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
G_{50}	1,115	479	613	1,120	662	**3,989**
G_{100}	279	114	129	104	153	**779**
G_{250}	334	131	148	113	188	**914**
G_{500}	189	77	86	37	118	**507**
G_{1000}	497	189	235	60	275	**1,256**
Total	**2,414**	**990**	**1,211**	**1,434**	**1,396**	**7,445**

위 표에서 리뷰 수가 100개 이하인 플레이스의 수, 즉 G_{50}과 G_{100}에 포함된 플레이스의 수는 총 4,768개로 전체의 64%를 차지하였다. 이들은 전체 플레이스에서 다수를 차지하므로 이들에 대한 리뷰 필터링이 제대로 이루어지지 않을 경우 “인근 플레이스 추천” 등을 포함한 다수의 소셜 빅데이터 분석 응용[21, 22]에서 추천 정확도가 크게 저하될 수 있다.

다음으로, G_b에 포함된 플레이스들의 리뷰 수를 총합한 분포를 확인하였다. 아래 표는 G_b별 리뷰 수의 분포를 보여준다. 한 가지 주목할 점은 G_{50}과 G_{100}에 포함된 플레이스들의 모든 리뷰를 합친 수가 11만여 개에 불과했으며, 전체 리뷰의 7.2% 정도를 차지하였다. 이는 전체의 64%를 차지하는 플레이스 수와 비교했을 때 대조적이었다.

Table 2. Number of reviews of the restaurants included in group G_b for each district of Ulsan

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
G_{50}	18,715	6,805	10,497	11,026	10,029	**57,072**
G_{100}	19,373	8,031	8,989	6,928	10,357	**53,678**
G_{250}	54,455	21,310	23,930	18,239	30,773	**148,707**
G_{500}	64,657	26,159	28,705	12,540	40,314	**172,375**
G_{1000}	441,722	163,229	210,792	52,790	239,013	**1,107,546**
Total	**598,922**	**225,534**	**282,913**	**101,523**	**330,486**	**1,539,378**

위와 같이 수집된 리뷰로부터 실험을 위해 각 구별로 5,000개의 데이터를 샘플링하였다. 지도 학습을 수행하기 위해서는 각 리뷰의 노이즈 여부를 수작업으로 라벨링해야 하며, 5,000이라는 값은 수작업 라벨링이 가능한 경험치로 설정되었다. 리뷰는 플레이스에 속해 있으므로, 데이터 샘플링 문제는 각 구별로 리뷰 수가 5,000개가 될 때까지 플레이스를 하나씩 선별하여 실험 데이터에 추가하는 문제로 귀결된다.

제안 방법에서는 플레이스를 선별할 때 Table 2에서 조사된 G_b별 리뷰 수의 비율을 고려하였다. 예를 들어, Table 2에서 울산 남구의 G_{50}, G_{100}, G_{250}, G_{500}, G_{1000}의 리뷰 수의 비율은 각각 3.1%, 3.2%, 9.1%, 10.8%, 73.8%였다. 이로부터 각각의 비율에 실험 데이터의 목표 크기인 5,000을 곱하여 G_b별 리뷰 수를 156, 162, 455, 540, 3,688로 계산해 낼 수 있다. 데이터 셋 구성을 위해 G_b별로 필요한 리뷰 수가 결정되면, G_b에 포함된 플레이스를 대상으로 목표한 리뷰 수가 만족될 때까지 랜덤 샘플링을 통해 플레이스를 추가한다.

Table 3는 위 방법을 통해 실험 데이터에 추가된 G_b별 플레이스 수를 보여준다. 그리고 Table 4는 실험 데이터에 추가된 G_b별 리뷰 수를 보여준다. 실험 데이터에는 총 114개의 플레이스를 대상으로 수집된 약 25,000개의 리뷰가 포함되었다. 울주를 제외한 나머지 지역구에서는 모두 5,000개 이상의 리뷰가 포함되었다. 울주의 경우, G_{500}과 G_{1000}에 속한 플레이스 수가 부족하여 선별된 리뷰 수가 5,000에 크게 못 미치는 것으로 확인되었다.

Table 3. Number of the restaurants included in the experimental data sets for each group G_b in Ulsan

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
G_{50}	7	8	9	24	7	**55**
G_{100}	2	3	2	3	2	**12**
G_{250}	4	3	3	5	3	**18**
G_{500}	2	1	2	0	3	**8**
G_{1000}	5	4	5	2	5	**21**
Total	**20**	**19**	**21**	**34**	**20**	**114**

Table 4. Number of reviews of the restaurants included in the experimental data sets for each group G_b in Ulsan

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
G_{50}	183	151	202	543	158	**1,237**
G_{100}	186	222	165	194	185	**952**
G_{250}	579	617	453	729	534	**2,912**
G_{500}	778	400	542	0	854	**2,574**
G_{1000}	4,156	4,345	4,380	1,169	4,127	**18,177**
Total	**5,882**	**5,735**	**5,742**	**2,635**	**5,858**	**25,852**

2. Algorithm Configuration

제안 방법에서 노이즈 리뷰 필터링은 기계 학습을 이용해 구현된다. 사용자 리뷰는 텍스트로 작성되어 있으므로, 여러 기계학습 알고리즘 중 텍스트 형식의 데이터 처리에 있어 높은 정확도를 제공하는 것으로 알려져 있는 LSTM과 BERT를 비교 대상으로 선정하였다.

LSTM은 RNN(Recurrent Neural Network)[23]과 함께 문장의 순서를 학습에 반영시킬 수 있어 자연어 처리에 높은 정확도를 제공하는 것으로 알려져 있다. RNN은 문장의 길이가 길어질 경우 정확도가 급격하게 떨어지는 문제가 있어 비교 대상에서 제외하였다. BERT는 트랜스포머를 기반으로 광범위한 데이터를 사전 훈련하여 만든 자연어 처리 전용 학습 모델로, 문장의 분류나 번역 등의 응용에서 높은 정확도를 제공한다. 범용적인 자연어 처리 응용에서는 BERT가 LSTM에 비해 높은 성능을 제공하는 것으로 알려져 있다.

따라서 본 논문에서는 노이즈 리뷰 필터링을 구현하기 위한 학습 알고리즘으로 LSTM과 BERT을 채택하고, 실험을 통해 성능을 비교하였다. 모델 구현에는 구글에서 제공하는 인공 신경망 오픈 소스 소프트웨어 라이브러리인 Keras[24]와 Python 언어가 이용되었다.

먼저 LSTM을 이용한 학습 모델은 아래와 같이 구현되었다. LSTM 모델은 Embedding, LSTM 및 2개의 Dense 층을 순차적으로 연결시킨 형태를 지닌다.

```
model = keras.Sequential([
    keras.layers.Embedding(2000, 50, input_length=40),
    keras.layers.LSTM(32),
    keras.layers.Dense(128, activation='relu'),
    keras.layers.Dense(1, activation='sigmoid')
])
```

Embedding 층은 리뷰에서 나타나는 식별 가능한 단어들의 집합인 Corpus 정보를 포함한다. 해당 층은 Corpus의 최대 크기, 단어의 차원, 문장의 최대 길이 등을 파라미터로 입력 받으며, 위 코드에서는 학습 데이터를 저장하는데 충분하도록 2000, 50, 40의 값을 차례로 입력하였다. 참고로 실험 데이터에서 각 리뷰는 평균 40개의 단어로 구성되어 있었으며, G_{1000}에 속한 플레이스들의 Corpus의 평균 크기는 1,900여개였다. 단어별 차원 수는 정확도 향상을 위해 조정될 수 있다.

LSTM 층에서는 Embedding 층으로부터 40×50 크기의 리뷰 정보를 전달받아 학습을 수행한다. 입력 파라미터 32는 학습 노드의 개수를 의미하며, 리뷰 별로 학습이 완료된 후 32 차원의 벡터가 다음 층으로 전달된다. 나머지 두 개의 Dense 층은 LSTM 층의 출력 결과로부터 리뷰의 노이즈 여부를 판단하기 위한 분류 목적으로 이용되며, 일반적인 심층 신경망 구조를 지닌다.

BERT를 이용한 학습 모델은 아래와 같이 구현되었다. 여러 버전의 사전학습 모델 중 실험에서는 다국어를 지원하는 12개의 층으로 구성된 기본 모델을 다운로드 받아 활용하였다.

```
bert = load_trained_model_from_checkpoint(config_path,
        model_path, ··· , seq_len=40)
outputs = keras.layers.Dense(1, activation='sigmoid')
        (bert.layers[-3].output)
model = keras.models.Model(bert.inputs, outputs)
```

다운로드된 사전학습 모델은 Keras의 load_trained_model_from_checkpoint() 함수를 이용하여 불러올 수 있으며, 해당 모델의 출력에 노이즈 여부 판별을 위한 Dense 층만 추가한 형태로 모델을 구성하였다. 위 코드에서 config_path와 model_path는 BERT 사전 모델의 저장 위치를 가리키며, seq_len에는 문장의 최대 길이를 지정한다. Dense 층의 입력은 BERT 모델의 출력 부분(위 코드에서 bert.layers[-3].output)과 연결된다. 최종 모델은 BERT 모델의 입력(bert.inputs)과 Dense 층의 출력(out puts)을 연결하여 완성된다.

IV. Experimental Results

실험에서는 Table 3과 4에서 제시된 샘플 데이터를 대상으로 LSTM과 BERT의 성능을 비교하였다. 실험은 Intel Xeon E5-2609 1.70GHz CPU와 MSI GeForce RTX-4090 GPU, 32GB 메모리가 장착된 HP 서버에서 수행되었다. 학습 알고리즘은 TensorFlow 2.10.1을 이용하여 구현되었으며, GPU 가속 라이브러리로 CUDA 11.8이 이용되었다.

1. Accuracy of Individual Learning

먼저 Table 3의 각각의 플레이스를 대상으로 LSTM과 BERT의 학습 정확도를 비교하였다. 플레이스 별로 수집된 리뷰를 80대 20의 비율로 나누고, 각각을 훈련과 테스트를 위해 이용하였다. 정확도를 판별하기 위한 기준으로는 f1-스코어가 이용되었다. F1-스코어는 정밀도(Precision)와 재현율(Recall)의 조화 평균 값에 해당하며, 클래스 데이터 간의 불균형을 반영한 정확도 값을 얻을 수 있다는 측면에서 채택되었다.

	G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}	Avg.
LSTM	0.245	0.425	0.894	0.902	0.926	**0.678**
BERT	0.647	0.792	0.925	0.956	0.960	**0.856**

Fig. 1. Average values of f1-scores shown in LSTM and BERT for the restaurants in Table 3

Fig. 1은 G_b 별로 LSTM과 BERT의 평균 f1-스코어를 보여준다. 실험 결과, BERT의 정확도가 LSTM에 비해 평균 17.8% 우수한 것으로 확인되었다. LSTM의 경우, G_{100} 이하의 식당을 대상으로 f1-스코어가 0.5 미만이었으며, 이는 리뷰가 100개 이하인 플레이스를 대상으로 LSTM이 활용되기 어려움을 의미한다. 리뷰가 100개 이상인 식당들의 평균 스코어 값은 0.889였다. BERT의 경우, 모든 G_b에서 0.5 이상의 값을 제공하였다. 특히, 리뷰가 100개 이상인 경우 평균값은 0.955로 매우 높은 정확도를 보였다. 따라서 개별 플레이스를 대상으로 한 노이즈 리뷰 필터링에는 BERT가 적합한 것으로 확인되었다.

다음으로 리뷰 수가 부족할 경우 LSTM과 BERT의 적용가능성 여부를 보다 구체적으로 판단하기 위해, Table 3의 각 플레이스를 대상으로 얻어진 f1-스코어와 Zero-R 값을 비교하였다. Zero-R이란 원본 데이터에서 가장 높은 빈도로 나타난 값의 비율로 계산될 수 있으며, 예측 정확도가 의미를 가지는지 판단하기 위한 기준으로 이용된다.

예를 들어, 어떤 플레이스의 일반 리뷰와 노이즈 리뷰의 비율이 각각 35%와 65%라면, Zero-R 값은 65%가 된다. 이 경우 학습을 통해 얻어진 분류기의 예측 정확도는 최소 65% 이상이 되어야 의미를 가진다고 할 수 있다.

	G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}	Total
# of places	55	12	18	8	21	**114**
LSTM	40	6	0	0	0	**46**
BERT	18	2	0	0	0	**20**

Fig. 2. Number of the restaurants in Table 3 whose f1-scores are less than their Zero-R values when using LSTM vs. BERT

Fig. 2는 Table 3의 각 플레이스에 대해 LSTM과 BERT를 적용했을 때, f1-스코어가 Zero-R보다 적은 플레이스의 수를 클래스 별로 합산하여 보여준다. LSTM의 경우, G_{100} 이하의 플레이스를 대상으로 전체의 40.4%에 해당하는 46개의 플레이스가 Zero-R 기준을 만족하지 못하는 것으로 확인되었다. 반면 BERT의 경우 G_{100} 이하에서 전체의 17.5%에 해당하는 20개의 플레이스가 기준을 만족하지 못했다. 이는 LSTM보다는 우수하지만 여전히 만족할만한 성능이라고는 볼 수 없다. 결과적으로 리뷰 수가 부족한 개별 플레이스를 대상으로 LSTM과 BERT 모두 만족할만한 노이즈 리뷰 필터링 정확도를 제공하지 못하는 것으로 조사되었다.

2. Accuracy of Group Learning

리뷰 수가 부족한 플레이스의 경우, 학습 대상이 되는 리뷰 수가 적어 높은 학습 정확도를 기대하기 어렵다. 이를 보완할 수 있는 방법 중 하나는 특성이 유사한 다른 플레이스들의 리뷰를 학습시킨 분류기(Classifier)를 리뷰 수가 부족한 플레이스의 노이즈 리뷰 필터링에 이용하는 것이다.

예를 들어, Table 3에서 울산 남구 식당의 경우, G_{1000}에 속한 5개 플레이스들의 리뷰 4,165개를 학습시켜 분류기를 얻을 수 있다. 그리고 해당 분류기는 동일한 그룹의 G_{50}이나 G_{100}에 속한 다른 플레이스들의 노이즈 리뷰 필터링에 이용될 수 있으며, 궁극적으로 Table 1에서 울산 남구에 포함된 2,414개의 모든 식당에 공통으로 적용될 수 있다. 본 논문에서는 이러한 학습 방법을 그룹 학습이라 지칭하며, 편의상 그룹의 학습 데이터는 Table 3와 4의 샘플 데이터에서 G_{1000}에 속한 플레이스들의 리뷰를 이용하는 것으로 가정한다.

개별 학습과는 달리, 그룹 학습의 경우 학습 데이터로 이용되는 리뷰에 대해 두 가지 전처리가 필요하다.

- 리뷰에 나타나는 플레이스의 이름을 모두 통일시켜야 한다. 예를 들어, 울산 남구 식당 그룹의 경우 G_{1000}에 속한 5개 플레이스의 리뷰들이 하나의 플레이스의 리뷰처럼 보이기 위해서는, 리뷰에서 나타나는 5개 플레이스 이름을 하나의 공통된 이름으로 변경시킬 필요가 있다.
- 학습 데이터에서 T회 이상 나타나는 단어들만 추출하여 Corpus에 포함시킨다. 그리고 학습 데이터 내 각각의 리뷰를 해당 단어들로만 표현한다. 이는 리뷰들의 유사도를 높이기 위함이다.

위 전처리 과정에서 T를 크게 설정한다고 해서 항상 높은 정확도를 얻을 수 있는 것은 아니다. 그 이유는 T를 크게 할수록 그룹 내 리뷰들의 유사도가 높아지는 반면, 각각의 리뷰에 포함되는 단어 수가 줄어 학습 데이터의 크기가 줄어들기 때문이다. 실험에서는 T를 5로 설정하였다. 5는 경험적인 값이며, 응용이나 데이터 특성에 따라 조정될 수 있다.

	G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}	Avg.
Individual	0.245	0.425	0.894	0.902	0.926	**0.678**
Group	0.803	0.935	0.937	0.971	0.970	**0.923**

Fig. 3. Average values of f1-scores shown in the individual and the group learning for the restaurants in Table 3 when using LSTM

실험에서는 먼저 개별 학습과 그룹 학습의 정확도를 비교하였다. Fig. 3은 LSTM을 이용할 경우 개별 학습과 그룹 학습의 f1-스코어 차이를 보여준다. 식당의 경우, G_{100} 이하에서 개별 학습과 그룹 학습의 평균값은 각각 0.335에서 0.869로, 그룹 학습을 이용할 경우 2.5배 이상 성능이 향상되었다. G_{250} 이상에서는 0.907에서 0.959로 5% 이상 성능이 향상되었다.

Fig. 4는 BERT를 이용할 경우 개별 학습과 그룹 학습의 f1-스코어 차이를 보여준다. 식당의 경우, G_{100} 이하에서 개별 학습과 그룹 학습의 평균값은 각각 0.719에서 0.873으로, 그룹 학습 이용 시 20% 이상 성능 향상이 발생하였다 G_{250} 이상에서는 0.947에서 0.959로 1% 정도 향상되었다. BERT는 이미 높은 개별 학습 정확도를 제공하고 있었으므로 LSTM에 비해 그룹 학습 도입에 따른 성능 향상이 그리 크지는 않는 것으로 확인되었다.

	G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}	Avg.
Individual	0.647	0.792	0.925	0.956	0.960	**0.856**
Group	0.812	0.934	0.931	0.977	0.971	**0.925**

Fig. 4. Average values of f1-scores shown in the individual and the group learning for the restaurants in Table 3 when using BERT

위 실험 결과에서 주목할 점은 그룹 학습 시 LSTM과 BERT의 정확도 차이가 거의 나지 않는다는 점이다. Fig. 5는 그룹 학습 시 LSTM과 BERT의 f1-스코어 차이를 보여준다. 두 알고리즘의 평균값은 각각 0.923과 0.925였으며, 차이가 0.2%에 불과하였다.

	G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}	Avg.
LSTM	0.803	0.935	0.937	0.971	0.970	**0.923**
BERT	0.812	0.934	0.931	0.977	0.971	**0.925**

Fig. 5. Average values of f1-scores shown in LSTM and BERT for the restaurants in Table 3 when using the group learning

다음으로 리뷰 수가 부족한 플레이스에 대해 그룹 학습이 효과적으로 적용될 수 있는지 확인하고자 하였다. Fig. 6은 개별 학습과 그룹 학습을 이용할 경우, f1-스코어가 Zero-R보다 적은 플레이스의 수를 G_b 별로 합산하여 보여준다. LSTM의 경우, 그룹 학습을 이용할 경우 f1-스코어가 Zero-R보다 적어 학습이 무의미한 플레이스의 수가 46에서 10으로 줄어들며, 전체 플레이스에서 차지하는 비율로 보았을 때 이는 기존의 40.4%에서 8.8%로 크게 줄어든 수치이다. BERT의 경우 역시 그룹 학습을 이용할 경우 20에서 10으로 줄어들었으며, 기존의 17.8%에서 8.8%로 줄어들었다. 두 알고리즘 모두 리뷰 수가 부족한 플레이스에 대해 그룹 학습을 적용할 경우 동일한 정확도를 보였으며, 개별 학습에 비해 정확도가 크게 개선됨을 확인할 수 있었다.

		G_{50}	G_{100}	G_{250}	G_{500}	G_{1000}
# of places		**55**	**12**	**18**	**8**	**21**
Individual	LSTM	40	6	0	0	0
	BERT	18	2	0	0	0
Group	LSTM	9	1	0	0	0
	BERT	8	1	1	0	0

Fig. 6. Number of the restaurants in Table 3 whose f1-scores are less than their Zero-R values when using the individual and the group learning

실험 결과, 그룹 학습의 경우 개별 학습 결과와는 달리 LSTM과 BERT의 정확도에 큰 차이가 없었다. 두 알고리즘의 f1-스코어 평균값은 0.923과 0.925로, 차이가 0.2% 이내였다. 그리고 리뷰 수가 부족한 플레이스에 대한 정확도 역시 동일한 결과를 제공하였다. 이에 반해, 실행 시간 측면에서는 [5]에서 언급한 바와 같이 LSTM이 BERT에 비해 더 나은 성능을 제공하는 것으로 알려져 있다. 따라서 노이즈 리뷰 필터링과 같이 응용 분야가 한정적일 경우, 사전 학습을 기반으로 한 BERT에 비해 비교적 간단한 형태의 LSTM이 보다 효과적으로 활용될 수 있음을 알 수 있다.

V. Conclusion and Future Work

본 논문에서는 리뷰 수가 부족한 플레이스를 대상으로 노이즈 리뷰 필터링의 정확도를 높이기 위한 지도 학습 방법을 소개하였다. 제안 방법은 유사한 특성을 가진 플레이스를 그룹으로 묶고, 이 중 일부 플레이스의 데이터를 학습시켜 분류기를 얻은 다음, 이를 그룹 내 다른 플레이스의 리뷰 필터링에 이용하는 방법이다. 본 논문에서는 그룹 학습과 관련한 초기 연구로써, 학습의 효율성과 정확도를 최대화하기 위한 구체적인 방법이나 이론을 제시하기 보다는, 온라인을 통해 수집된 실제 데이터를 활용한 실험을 통해 그룹 학습의 적용 가능성을 확인하고자 하였다.

실험을 위해 울산광역시에 등록된 16,900여 개의 식당을 대상으로 약 154만개의 리뷰를 수집하였으며, 이 중 114개의 플레이스의 25,000여개의 리뷰를 샘플링하여 지도 학습을 수행하였다. 학습 알고리즘은 텍스트 처리에 우수한 성능을 보이는 LSTM과 BERT를 이용하였다. 실험을 통해 아래와 같은 결과를 확인하였다.

- 각각의 플레이스를 대상으로 개별 학습을 수행하는 기존 방법에 비해, 제안한 그룹 학습의 정확도가 우수하였다. LSTM의 경우, 개별 학습과 그룹 학습의 정확도(f1-스코어)는 각각 0.678과 0.923으로 조사되었으며(Fig. 3 참조), BERT의 경우는 각각 0.856과 0.925였다(Fig. 4 참조).
- 그룹 학습은 리뷰 수가 적어 학습이 어려운 플레이스를 대상으로도 효과적으로 적용될 수 있었다. 리뷰 수가 100개 미만인 경우, 개별 학습이 어려운 플레이스의 비율은 LSTM과 BERT에서 각각 40.4%, 17.8%였으며, 그룹 학습을 적용할 경우 비율이 동일하게 8.8%로 줄어들었다.

위 실험 결과를 통해, 제안한 그룹 학습이 리뷰 수가 부족한 플레이스를 대상으로 한 노이즈 리뷰 필터링에 효과적으로 적용될 수 있음을 확인하였다. 실험 결과에서 한 가지 주목할 점은 그룹 학습의 경우 LSTM과 BERT의 정확도가 0.2% 이내였다는 점이다. 따라서 컴퓨팅 자원이 부족할 경우 LSTM이 보다 효과적으로 이용될 수 있음을 확인하였다.

향후에는 그룹 학습의 효율성과 정확도를 최대화하기 위한 플레이스 그룹핑 기준과 학습 데이터 샘플링 방법에 대해 연구할 계획이다. 또한 리뷰수가 부족한 플레이스에 대해 정확도를 더욱 높일 수 있는 방안에 대해서도 함께 연구를 지속할 예정이다.

REFERENCES

[1] K. H. Lee et al., "Parallel Data Processing with Map Reduce: a Survey," ACM SIGMOD Record, Vol. 40, No. 4, pp. 11-20, 2012.

[2] E. F. Cardoso, R. M. Silva, and T. A. Almeida, “Towards automatic filtering of fake reviews,” Neurocomputing, vol. 309, pp. 106-116, 2018. DOI: 10.1016/j.neucom.2018.04.074

[3] Naver Open API, https://developers.naver.com/docs/common/openapiguide/

[4] Google Developer API, https://developers.google.com/

[5] H. G. Kim, “Efficient filtering of noise reviews using supervised learning in social big data analysis,” Journal of the Korea Society of Computer and Information, Vol. 28, No. 6, 2023.

[6] S. Hochreiterand and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.

[7] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[8] N. Jindal and B. Liu, “Review spam detection,” in Proc. of the 16st international conference on World Wide Web (WWW), pp. 1189-1190, May 2007. DOI: 10.1145/ 1242572.1242759

[9] H. Li, Z. Chen, A. Mukherjee, B. Liu, and J. Shao, “Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns,” in Proc. of the 9th International AAAI Conference on Web and Social Media (ICWSM), pp. 634-637, 2015. DOI: 10.1609/ icwsm.v9i1.14652

[10] C. L. Lai, K. Q. Xu, R. Y. K. Lau, Y. Li, and L. Jing, “Toward a language modeling approach for consumer review spam detection,” in Proc. of the IEEE International Conference on E-Business Engineering, Shanghai, China, pp. 1-8, 2010. DOI: 10.1109/icebe.2010.47

[11] A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in Proc. of the 21st international conference on World Wide Web (WWW), pp. 191-200, 2012. DOI: 10.1145/2187836.2187863

[12] A. Barushka, and P. Hajek, “Review spam detection using word embeddings and deep neural networks,” in Proc. of the 15th IFIP WG 12.5 International Conference, AIAI 2019, Hersonissos, Crete, Greece, pp. 340-350, 2019. DOI: 10.1007/978-3-030-19823-7_28

[13] G. M. Shahariar et al., “Spam review detection using deep learning,” in Proc. of the IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp.27-33, 2019. DOI: 10.1109/iemcon.2019.8936148

[14] K. O’Shea and R. Nash, “An introduction to convolutional neural network,” arXiv preprint arXiv:1511.08458, 2015.

[15] P. Gupta, S. Gandhi, and B. R. Chakravarthi, “Leveraging transfer learning techniques-bert, roberta, albert and distilbert for fake review detection,” in Forum for Information Retrieval Evaluation, pp. 75-82, 2021.

[16] Y. Shang, M. Liu, T. Zhao, and J. Zhou, “T-bert: A spam review detection model combining group intelligence and personalized sentiment information,” in Proc. of 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, pp. 14-17, September 2021. DOI: 10.1007/978-3-030-86383-8_33

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[18] G. Bathla, P. Singh, R K. Singh, E. Cambria, and R. Tiwari, “Intelligent fake reviews detection based on aspect extraction and analysis using deep learning,” Neural Comput. Appl., vol. 34, pp. 20213-20229, 2022. DOI: 10.1007/s00521-022-07531-8

[19] P. Bhuvaneshwari, A. M. Rao, and Y. H. Robinson, “Spam review detection using self-attention based CNN and bi-directional LSTM,” Multimed. Tools. Appl., vol. 80, pp.18107-18124, 2021. DOI: 10.1007/s11042-021-10602-y

[20] Small Business Promotion Agency, http://sg.sbiz.or.kr/

[21] Y. W. Yu, S. H. Kim, and H. G. Kim, “Travel route recommendation utilizing social big data,” Journal of the Korea Society of Computer and Information, Vol. 27, No. 5, pp. 117-125, 2022.

[22] H. G. Kim, “Implementation of a travel route recom- mendation system utilizing daily scheduling templates,” Journal of the Korea Society of Computer and Information, Vol. 27, No. 10, pp. 137-146, 2022.

[23] L. R. Medsker and L. C. Jain, “Recurrent neural networks,” Design and Applications, vol. 5, pp. 64-67, 2001.

[24] Google Keras: https://www.tensorflow.org/guide/keras

Authors

Hyeon Gyu Kim received the B.S. and M.S. degrees in Computer Science from University of Ulsan, and Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology, Korea, in 1997, 2000, and 2010, respectively. Dr. Kim joined the faculty of the Division of Computer Science and Engineering at Sahmyook University, Seoul, Korea, in 2012. He is currently an Associate Professor in the Division of Computer Science and Engineering, Sahmyook University. He is interested in artificial intelligence and big data processing.