JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 6, pp. 63-69, June 2023
https://doi.org/10.9708/jksci.2023.28.06.063

Efficient Filtering of Noise Reviews Using Supervised Learning in Social Big Data Analysis

Hyeon Gyu Kim*

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

[Abstract]

Social reviews collected through the search API may include a large number of reviews unrelated to a given search term, and these reviews are referred to as noise reviews because they may lead to distorted analysis results. In this paper, we discuss supervised learning algorithms to conduct filtering of the noise reviews efficiently, and compare their performance through experiments. About 20,000 reviews collected for tourist attractions in the Ulsan metropolitan city were used for the experiments, and LSTM and BERT, which are known to provide high accuracy in text processing, were adopted for training and testing the reviews. As a result, BERT provided better accuracy than LSTM, where f1-scores of the two algorithms were 90.1% and 95.2%, respectively. On the other hand, in terms of execution time, LSTM was about 5 times faster than BERT. The result shows that, in the noise review filtering, BERT can be used more properly when accuracy is important, whereas LSTM can be used more properly when performance is important or computation resources are insufficient.

▸ **Key words**: Social big data, Noise review filtering, Supervised learning, LSTM, BERT

[요 약]

검색 API를 통해 수집된 소셜 리뷰에는 주어진 검색어와 상관없는 리뷰가 다수 포함되어 있을 수 있으며, 이들 리뷰는 왜곡된 분석 결과를 초래할 수 있어 노이즈 리뷰로 지칭된다. 본 논문에서는 노이즈 리뷰를 효과적으로 필터링하기 위한 지도 학습 알고리즘들을 소개하고, 실험을 통해 이들의 성능을 비교한다. 실험에는 울산광역시의 관광지를 대상으로 수집된 2만여 개의 리뷰가 이용되었으며, 학습 알고리즘으로는 텍스트 처리에 높은 정확도를 제공하는 것으로 알려진 LSTM과 BERT를 이용하였다. 실험 결과, BERT의 정확도가 LSTM에 비해 우수했으며, 두 알고리즘의 f1-스코어는 각각 90.1%와 95.2%로 조사되었다. 반면, 실행시간 측면에서는 LSTM이 BERT에 비해 5배 정도 빠른 성능을 제공하였다. 따라서 노이즈 리뷰 필터링 문제에 있어 정확도가 중요한 경우 BERT가, 성능이 중요하거나 컴퓨팅 자원이 부족한 경우는 LSTM이 보다 적절하게 이용될 수 있다.

▸ **주제어**: 소셜 빅데이터, 노이즈 리뷰 필터링, 지도 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. 05. 19, Revised: 2023. 06. 07, Accepted: 2023. 06. 08.

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

I. Introduction

빅데이터란 3V(Volume, Velocity, Variety) 속성을 지니는 데이터로 정의될 수 있으며, 대표적인 사례로 신용카드 트랜잭션, 휴대폰 전화 및 문자 내역, SNS 피드 및 블로그 리뷰 등을 포함한다[1, 2]. 이 중, SNS 피드 및 블로그 리뷰 등을 포함한 소셜 빅데이터는 고객 관점의 의견이나 불만 사항을 추출하기 위해 대중적으로 이용되고 있다. 소셜 빅데이터는 온라인 포털 업체 등이 제공하는 오픈 API[3, 4] 등을 통해 무료로 획득 가능하며, 텍스트 형태로 구성되어 (신용카드, 휴대폰 내역 등 수치로 이루어진 데이터에 비해) 의견이나 불만 사항의 직접적인 원인을 바로 알아낼 수 있다는 점에서 선호된다.

오픈 API를 이용하여 수집된 리뷰에는 주어진 검색어와 연관성이 떨어지는 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 관광지인 “대왕암”이 주어졌을 때, 아래와 같은 리뷰가 검색 결과에 포함될 수 있다.

- 울산 대왕암 맛집 보리꽃 푸짐해요 …
- 깔끔하고 넓은 매장의 대왕암공원 인근 맛집 하라검 소개합니다 …
- 울산 대왕암 분위기 카페 하몬 인스타 감성 낭낭한 곳

문제는 이들 리뷰가 검색 결과에서 차지하는 비율이 너무 높다는 점이다. 관광지의 경우 약 72%의 리뷰가 주어진 검색어와 연관성이 떨어지는 것으로 조사되었다 (4장 참조). 따라서 위 시나리오에서 정확한 결과를 얻기 위해서는 이들 리뷰를 필터링하기 위한 작업이 반드시 선행되어야 한다. 이들 리뷰가 분석 결과를 왜곡한다는 측면에서 본 논문에서는 편의상 이들 리뷰를 “노이즈 리뷰”로 지칭한다.

노이즈 리뷰는 다수의 기존 연구에서 소개한 스팸 리뷰와는 다소 차이가 있다. 스팸 리뷰는 주어진 상품이나 플레이스에 대한 허위 및 과장 광고를 통해 광고 효과를 극대화하기 위한 목적으로 작성되며, 주어진 상품에 대한 정보를 직접적으로 다루는 것으로 볼 수 있다. 반면 노이즈 리뷰는 다른 상품을 소개하는 과정에서 주어진 상품이 간접적으로 언급되는 형태이다. 따라서 스팸 리뷰 필터링의 정확도를 높이기 위해서는 노이즈 리뷰를 걸러내는 작업이 선행될 필요가 있다. 그러나 그 중요성에도 불구하고, 노이즈 리뷰 필터링 문제를 다룬 연구는 현재까지 찾아보기 어려운 실정이다.

본 논문에서는 노이즈 리뷰를 효과적으로 필터링하기 위한 지도 학습 알고리즘들을 소개하고, 실험을 통해 이들의 성능을 비교한다. 실험을 위해 울산광역시의 45군데의 관광지를 대상으로 수집된 2만 여개의 소셜 리뷰에 대한 노이즈 여부를 라벨링하고 학습시켰다. 학습 알고리즘으로는 텍스트 처리에 높은 정확도를 제공하는 것으로 알려진 LSTM(Long Short-Term Memory)[5]과 BERT(Bidirectional Encoder Representation for Transformers)[6]가 이용되었다.

본 논문의 구성은 다음과 같다. 2장에서는 스팸 리뷰 필터링과 관련한 기존 연구들에 대해 소개한다. 3장에서는 노이즈 리뷰 필터링을 구현하기 위한 LSTM과 BERT의 학습 모델에 대해 설명한다. 4장에서는 학습 정확도와 실행 시간 측면에서 두 학습 모델의 성능을 실험하고 결과를 비교한다. 5장에서는 결론 및 추후 연구 방향 제시로 마무리한다.

II. Related Work

[7]에서는 스팸 리뷰를 (i) 광고, (ii) 특정 브랜드를 분석하는 리뷰, (iii) 특정 제품을 홍보하거나 훼손할 목적으로 쓰여진 리뷰 등으로 분류했다. 이 중 첫 번째와 두 번째 형태의 리뷰는 전통적인 기계 학습으로 비교적 쉽게 해결될 수 있는데 반해, 마지막 형태의 리뷰는 사용자들을 기만하기 위해 리뷰가 진짜처럼 보이도록 신중하게 작성되었기 때문에 이들을 필터링하기 위해서는 좀 더 복잡한 형태의 솔루션이 필요하다고 언급하였다.

[8]에서는 사용자 기만 목적의 리뷰들을 필터링하기 위해, 텍스트 컨텐츠를 기반으로 한 전통적인 분류 기법 이외에도, 행위나 그룹을 기반으로 한 분류 기법이 함께 이용될 수 있다고 소개하였다. 이와 관련하여, [9]에서는 리뷰의 포스팅 시간이나 위치 등을 기반으로 리뷰를 작성한 사용자의 동적 행위를 분석하여 필터링에 이용하기 위한 방법을 제시하였다. [10]에서는 판매량, 가격, 별점 등의 상품 관련 정보를 활용하여 필터링의 정확도를 높이기 위한 방법을 논의하였으며, [11]에서는 스팸을 보내는 사용자들을 그룹핑하고 이들의 유사성을 추출하여 스팸 리뷰 분류에 활용하였다.

최근 들어, 스팸 리뷰 필터링 문제에 딥러닝 기법을 적용하여 정확도를 높이고자 하는 많은 시도가 이루어졌다. [12]는 Word2Vec 등의 단어 임베딩 기법과 심층 신경망(DNN, Deep Neural Network)을 이용하여 스팸 리뷰 필

터링을 구현하였으며, 89%의 정확도를 제공하였다. [13]에서는 CNN과 LSTM을 이용하여 필터링을 구현하였으며, 93.6%의 정확도를 제공하였다. [14] 역시 CNN과 LSTM을 이용하여 스팸 리뷰를 필터링하였으며, 최대 95.5%의 정확도를 제공하였다. 이들 연구에서는 공통적으로 딥러닝을 이용할 경우 별도의 행위나 부가 정보 없이 텍스트 분석만으로 고도의 필터링 정확도를 얻을 수 있음을 실험을 통해 확인하였다.

단, 위에서 언급된 기존 연구들이 노이즈 리뷰 필터링에도 적용될 수 있을지는 미지수이다. 이와 관련하여, 본 논문에서는 노이즈 리뷰 필터링 문제에 딥러닝 기법을 적용해 높은 정확도를 얻을 수 있을지와 관련하여 실험을 통해 확인하고자 하였다.

III. Algorithms for Review Filtering

사용자 리뷰는 자연어로 쓰여져 있다. 자연어로 작성된 문장에서는 동일한 단어라도 문장 내의 위치나 함께 이용되는 단어의 종류에 따라 의미가 달라질 수 있다. 따라서 자연어 문장의 정확한 의미를 파악하기 위해서는 문장에서 나타나는 단어의 종류와 순서를 함께 고려해야 한다. 이로부터 데이터의 순서를 고려한 학습 방법인 RNN (Recurrent Neural Network)[15]이나 LSTM 등이 자연어 처리에서 주로 활용되어 왔다.

RNN과 LSTM은 신경망(Neural Network)을 기반으로 하며, 망을 구성하는 각각의 Hidden 노드에서 순서를 기억하기 위한 추가 메모리를 도입하여 이용한다는 점에서 공통점을 지닌다. 입력 데이터는 순서를 지니는 시퀀스의 형태로 구성되며, 두 알고리즘에서는 입력 시퀀스의 길이에 따라 성능의 차이가 발생한다. 보다 자세히, RNN은 입력 시퀀스의 길이가 길어질 경우 기울기 소실(Vanishing Gradient) 문제가 발생하여 정확도가 급격하게 떨어지는 단점이 있으며, LSTM은 Hidden 노드에 Input, Forget, Output 게이트 등의 추가적인 메모리를 도입해 기울기 소실 문제를 해결하고자 하였다. 일반적으로 LSTM이 RNN에 비해 우수한 성능을 제공하는 것으로 알려져 있다.

최근에는 트랜스포머(Transformer)를 기반으로 광범위한 데이터를 사전 훈련하여 만든 BERT가 발표되었으며, 자연어 처리에 높은 성능을 제공하는 것으로 알려져 있다. BERT는 사전 훈련 모델이므로 전이 학습(Transfer Learning) 형태로 이용될 수 있다. 즉, 사전 학습된 Bert 모델을 다운받아 이를 기반으로 DNN이나 LSTM 층을 추가시킨 형태로 모델 구현이 가능하다. 대량의 Corpus를 이용해 사전 훈련되어 있기 때문에 기본적인 형태의 DNN만 추가해도 우수한 성능이 제공되는 것으로 알려져 있다.

본 논문에서는 노이즈 리뷰 필터링을 구현하기 위한 학습 알고리즘으로 LSTM과 BERT을 채택하고, 실험을 통해 성능을 비교하였다. 앞서 언급한 바와 같이, RNN은 문장이 길어질 경우 정확도가 급격하게 떨어지는 문제가 있어 비교 대상에서 제외하였다. 모델 구현에는 구글에서 제공하는 인공 신경망 오픈 소스 소프트웨어 라이브러리인 Keras[16]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. Corpus 내 각각의 단어는 다차원 벡터 형태로 표현된다. 이는 단어 별로 순서나 다른 단어와의 관계 정보를 유지하기 위함이다. Embedding 층은 Corpus의 최대 크기, 단어의 차원, 문장의 최대 길이 등을 파라미터로 입력 받는다. 위 코드에서는 학습 데이터를 저장하는데 충분하도록 2000, 50, 40의 값을 차례로 입력하였다. 이 경우, 각 단어는 50차원의 벡터로 표현되며, 다음 단계인 LSTM 층으로 각각의 리뷰 정보가 40×50 크기의 행렬 형태로 전달된다. 참고로 실험 데이터에서 관광지 별 평균 리뷰 수는 520개였으며, 각 리뷰는 평균 40개의 단어로 구성되어 있었다. 그리고 관광지 별 Corpus의 평균 크기는 1,700여개였다.

LSTM 층에서는 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 신경망 구조를 지닌다. 첫 번째 Dense 층은 128개의 학습 노드를 가지며, 분류에 필요한 가중치 정보들을 노드에 저장한다. 이들 정보는

Relu 활성화 함수를 거쳐 다음 층으로 전달된다. 두 번째 Dense 층은 전달된 가중치 값을 합산하고 sigmoid 함수를 적용하여 노이즈 여부 판별을 위한 최종 값을 출력한다.

LSTM은 주어진 데이터만을 학습에 이용한다. 따라서 학습에 필요한 파라미터의 크기가 BERT에 비해 상대적으로 작다. 예를 들어, LSTM의 Embedding 층에서 학습에 이용되는 파라미터 개수는 corpus의 크기와 단어의 차원 수를 곱한 값으로 계산될 수 있다. 위 구현 예의 경우, 100,000(= 2000 × 50)개의 파라미터가 Embedding 층에서 이용된다. 이에 반해 BERT에서는 주어진 데이터 외에 사전 학습된 데이터도 함께 이용된다. 예를 들어 BERT의 다국어 지원 기본 모델을 이용할 경우, Embedding 층에서 학습에 이용되는 파라미터 수는 약 91,000,000개 정도가 된다. 이는 BERT가 Corpus에 더욱 풍부한 정보를 포함하고 있음을 의미하며, 상대적으로 높은 학습 정확도를 제공하는 기초가 된다.

```
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)
```

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

IV. Experimental Results

본 논문에서는 울산광역시에 등록된 관광지를 대상으로 실험을 진행하였다. 울산광역시는 남구, 동구, 북구, 울주, 중구 등의 5개의 지역구로 구성되어 있으며, 7개 광역시 중 규모가 가장 작아 실험을 위해 수집해야 할 데이터의 크기가 비교적 작다는 점에서 실험 대상으로 선정되었다. 관광지 정보는 한국관광공사에서 제공하는 API[17]를 이용하여 얻을 수 있으며, 이를 통해 45개의 관광지 정보를 추출하였다.

수집된 각각의 관광지를 대상으로 네이버에서 제공하는 검색 API[3]를 이용하여 사용자 리뷰 수집 작업이 진행되었다. 리뷰 수집은 2023년 1월에 수행되었으며, 45개의 관광지를 대상으로 약 2만 여개의 리뷰가 수집되었다. 수집된 각각의 리뷰에는 지도 학습을 수행하기 위한 노이즈 리뷰 여부를 표시하였으며, 라벨링 작업은 수작업으로 진행되었다. Table 1은 울산광역시의 각 구별로 수집된 관광지와 리뷰 수를 함께 보여준다.

Table 1. Number of tourist attractions and their reviews for each district of the Ulsan metroporitan city

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
# of places	7	9	10	8	11	**45**
# of reviews	3,844	3,968	4,003	4,697	4,326	**20,838**

Table 2는 각 구별로 수집된 리뷰에서 노이즈 리뷰의 수와 비율을 보여준다. 관광지를 대상으로 수집된 전체 리뷰 중 72.1%가 노이즈 리뷰인 것으로 조사되었으며, 이는 소셜 리뷰 분석에 있어 높은 정확도를 얻기 위해서는 노이즈 리뷰를 필터링이 반드시 수행되어야 함을 나타낸다.

Table 2. Ratio of the noise and normal reviews collected for the tourist attractions in Table 1

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
# of noise reviews	2,740	2,798	2,804	3,597	3,062	**15,029**
# of normal reviews	1,104	1,170	1.199	1,100	1,264	**5,809**
Ratio of the noise reviews	**71.3%**	**70.5%**	**70.1%**	**76.6%**	**70.8%**	**72.1%**

실험에서는 수집된 관광지에 대해 리뷰 수에 따라 4개의 그룹으로 분류하였다. 이는 리뷰 수에 따른 학습 정확도 차이를 확인하기 위함이다. 네이버 검색 API를 이용할 경우 플레이스 별로 매월 수집될 수 있는 리뷰의 최대 수는 1,000개이다. 이로부터 1000, 500, 250, 100을 기준으로 리뷰 수 구간을 4개로 나누고, 각 플레이스를 리뷰 수에 따라 해당 구간의 그룹으로 분류할 수 있다. 관광지 집합을 P라고 하고, i번째 관광지를 p_i라고 하자. p_i의 리뷰 수를 라고 n_i하면, p_i는 그룹 G_b에 아래와 같이 할당될 수 있다.

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

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 그룹 G_b에 포함된 관광지 수를 보여준다.

Table 3. Number of the tourist attractions included in group G_b for each district of the Ulsan metropolitan city

	Namgu	Dongu	Bukgu	Ulju	Joongu	Total
G_{100}	0	0	0	1	0	**1**
G_{250}	0	1	1	1	2	**5**
G_{500}	2	1	1	0	1	**5**
G_{1000}	5	7	8	6	8	**34**
Total	**7**	**9**	**10**	**8**	**11**	**45**

실험에서는 각각의 관광지를 대상으로 지도 학습을 수행하고, LSTM과 BERT의 성능을 비교하였다. 플레이스별로 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 불균형이 있을 때 이를 반영한 정확도 값을 산정하기 위한 목적으로 활용된다. Table 2에서 볼 수 있듯이, 노이즈 리뷰의 비율이 70% 이상이므로 데이터 불균형을 고려하여 정확도를 측정할 필요가 있다.

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.

	Namgu	Dongu	Bukgu	Ulju	Joongu	Avg.
G_{100}	-	-	-	0.857	-	**0.857**
G_{250}	-	0.923	0.885	0.5	0.821	**0.79**
G_{500}	0.889	0.923	0.909	-	0.894	**0.901**
G_{1000}	0.926	0.917	0.933	0.913	0.907	**0.919**
Avg.	**0.915**	**0.919**	**0.926**	**0.846**	**0.89**	**0.901**

Fig. 1. Average values of f1-scores shown in LSTM for the tourist attractions of group Gb in each district of the Ulsan metropolitan city

	Namgu	Dongu	Bukgu	Ulju	Joongu	Avg.
G_{100}	-	-	-	0.833	-	**0.833**
G_{250}	-	1	0.967	1	0.935	**0.967**
G_{500}	0.956	0.98	0.971	-	0.959	**0.968**
G_{1000}	0.944	0.959	0.963	0.944	0.939	**0.951**
Avg.	**0.948**	**0.966**	**0.964**	**0.936**	**0.94**	**0.952**

Fig. 2. Average values of f1-scores shown in BERT for the tourist attractions of group Gb in each district of the Ulsan metropolitan city

Fig. 1은 관광지 그룹별로 LSTM의 f1-스코어의 평균값을 보여준다. 전체 평균값은 0.901이었으며, G_b별 평균값으로부터 리뷰 수가 증가할수록 정확도가 높아지는 것을 확인할 수 있었다.

Fig. 2는 관광지 그룹별로 BERT의 f1-스코어의 평균값을 보여준다. 전체 평균값은 0.952였으며, LSTM에 비해 약 5%정도 높은 정확도를 제공하였다. 한 가지 특이한 점은 BERT의 경우 리뷰 수에 따라 정확도가 비례하지 않고 일정한 수준으로 유지되었다. 이는 리뷰 수가 부족하더라도 사전 학습된 데이터를 활용함으로써 높은 정확도를 유지할 수 있는 것으로 파악되었다. 단, 리뷰수가 100 이하인 경우에는 LSTM과 유사하게 80%대의 정확도를 제공하였다.

단, 3장에서 언급한 바와 같이, BERT는 학습해야 할 파라미터가 많으므로 LSTM에 비해 더 많은 컴퓨팅 자원과 실행 시간을 요구할 수 있다. Fig. 3은 LSTM과 BERT의 실행 시간을 보여준다. 실험 결과, LSTM의 실행 속도가 BERT에 비해 5배 정도 빨랐다. GPU를 이용하지 않고 CPU만 이용할 경우 BERT의 수행 시간은 1시간을 넘어가는 경우가 많아 확인이 어려웠다. 따라서 컴퓨팅 자원이 부족한 경우에는 LSTM이 노이즈 리뷰 필터링에 더 적합하게 이용될 수 있음을 확인하였다.

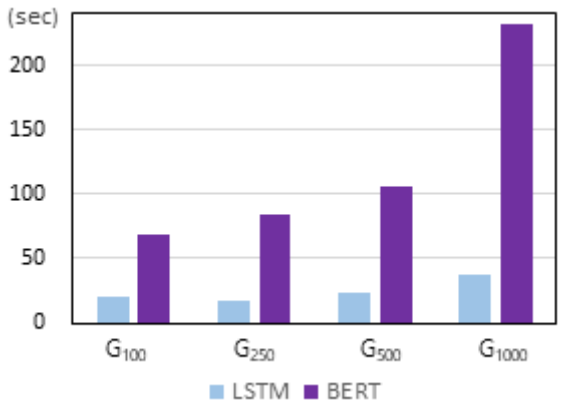

Fig. 3. Average execution time of LSTM and BERT for the tourist attractions of group G_b in each district of the Ulsan metropolitan city

V. Conclusion and Future Work

본 논문에서는 온라인에서 수집된 소셜 리뷰로부터 주어진 검색어와 연관성이 떨어지는 노이즈 리뷰를 효과적으로 필터링하기 위한 지도 학습 알고리즘들을 소개하고, 실험을 통해 이들의 성능을 비교하였다. 실험을 위해 울산광역시의 관광지를 대상으로 수집된 2만 여개의 리뷰에 대한 노이즈 여부를 라벨링하고 학습시켰으며, 학습 알고리즘으로는 텍스트 처리에 높은 정확도를 제공하는 것으로 알려진 LSTM과 BERT를 채택하였다. 실험 결과, LSTM과 BERT의 f1-스코어는 각각 0.901과 0.952로 조사되었으며, BERT의 정확도가 LSTM에 비해 평균 5% 정도 우수한 것으로 나타났다. LSTM에서는 리뷰 수에 따라 학습 정확도가 비례하는 반면, BERT에서는 리뷰 수에 관계없이 정확도가 일정하게 유지되었다. 이는 주어진 학습 데이터 외에도 사전 학습한 내용을 리뷰 필터링에 함께 적용함으로써, 학습 정확도가 입력 데이터의 크기에 비교적 영향을 덜 받는 것으로 확인되었다. 반면 BERT의 경우 학습해야 할 파라미터가 많아 LSTM에 비해 5배 정도 많은 학습 시간이 소요되었다. 따라서 컴퓨팅 자원이 부족한 경우, LSTM이 더 적합한 것으로 조사되었다.

본 논문에서는 관광지를 대상으로 노이즈 리뷰 필터링 관련 실험을 수행하였다. 관광지의 경우, 리뷰가 상대적으로 풍부하여 지도 학습을 수행하는데 큰 어려움이 없었다. 이에 반해 식당이나 카페의 경우 리뷰가 부족할 수 있으며, 학습 데이터가 부족할 경우 지도 학습을 수행하는데 어려움이 있을 수 있다. 따라서 향후에는 다양한 형태의 데이터를 수집하여 특성을 분석하고, 리뷰 수가 부족한 경우 정확도를 높이는 방안과 관련하여 연구가 필요한 상황이다.

또한 본 논문에서는 노이즈 리뷰 필터링 문제를 해결하기 위한 초기 접근 방법으로, 실험을 통해 기계 학습을 이용한 해결 가능성을 확인하는데 목적이 있었다. 학습 성능이나 정확도를 최대화하기 위한 구체적인 방법이나 이론적인 체계에 대해서는 다루지 않았다. 따라서 향후 연구에서는 노이즈 리뷰 필터링 문제와 관련한 이론적인 체계에 대해 연구를 지속할 예정이다.

REFERENCES

[1] W. L. Kang, H. G. Kim, and Y, J. Lee, "Reducing IO Cost in OLAP Query Processing with MapReduce," IEICE Trans. Inf. & Syst, Vol. E98-D, No. 2, pp. 444-447, Feb. 2015. DOI: 10.1587/transinf.2014edl8143

[2] K. H. Lee et al., "Parallel Data Processing with Map Reduce: a Survey," ACM SIGMOD Record, Vol. 40, No. 4, pp. 11-20, December 2011.

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

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

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

[6] 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.

[7] 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

[8] E. F. Cardoso, R. M. Silva, and T. A. Almeida, “Towards automatic filtering of fake reviews,” Neurocomputing, vol. 309, pp. 106-116, May 2018. DOI: 10.1016/j.neucom.2018. 04.074

[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] 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, May 2021. DOI: 10.1007/s11042-021-10602 -y

[14] 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, July 2022. DOI: 10.1007/s00521-022-07531-8

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

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

[17] Tour API 3.0, Korea Tourism Organization, https://kto.visitkorea.or.kr/kor/gov30/tourapi.kto

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.