딥러닝 모델의 정확도 향상을 위한 감성사전 기반 대용량 학습데이터 구축 방안

최민성○, 박상민, 온병원*
군산대학교, 소프트웨어융합공학과
alstjd517@kunsan.ac.kr, b1162@kunsan.ac.kr, bwon@kunsan.ac.kr

A Method of Constructing Large-Scale Train Set Based on Sentiment Lexicon for Improving the Accuracy of Deep Learning Model

Min-Seong Choi○, Sang-Min Park, Byung-Won On*
Department of Software Convergence Engineering, Kunsan National University

요 약

감성분석(Sentiment Analysis)은 텍스트에 나타난 감성을 분석하는 기술로 자연어 처리 분야 중 하나이다. 한국어 텍스트를 감성분석하기 위해 다양한 기계학습 기법이 많이 연구되어 왔으며 최근 딥러닝의 발달로 딥러닝 기법을 이용한 감성분석도 활발해지고 있다. 딥러닝을 이용해 감성분석을 수행할 경우 좋은 성능을 얻기 위해서는 충분한 양의 학습데이터가 필요하다. 하지만 감성분석에 적합한 학습데이터를 얻는 것은 쉽지 않다. 본 논문에서는 이와 같은 문제를 해결하기 위해 기존에 구축되어 있는 감성사전을 활용한 대용량 학습데이터 구축 방안을 제안한다.

주제어: 감성분석, 감성사전, 딥러닝, 학습데이터

1. 서론

스마트 기기, 인터넷의 발달로 인해 많은 양의 텍스트가 생성되고 있다. 이에 따라 자연어 처리와 함께 비정형 데이터인 텍스트의 감성을 분석하는 연구도 많이 증가하고 있는 추세이다. 감성분석이란 텍스트에 나타난 사람들의 태도, 의견, 성향과 같은 주관적인 데이터를 분석하는 자연어 처리 기술이다[1]. 생성된 텍스트에서 감성을 자동으로 분석할 수 있다면 가치 있고 유용한 정보를 얻을 수 있으며 많은 분야에서 활용이 가능하다. 특히 감성분석은 여론 분석이나 마케팅 도구로도 많이 활용되어 중요성이 점차 높아지고 있다. 최근 들어 딥러닝의 발달로 인해 딥러닝을 이용한 감성분석도 많이 연구되고 있다. 딥러닝은 학습데이터(Train Set)를 사용하여 모델을 학습, 구축하고 이와 같은 모델을 통해 평가데이터(Test Set) 분류하는 기법이다. 딥러닝 기법은 충분한 양의 학습데이터로 모델을 학습해야 좋은 모델을 구축할 수 있다. 하지만 감성분석에 적합한 충분한 양의 학습데이터를 구하기란 쉽지 않다.

본 논문에서는 이와 같은 문제를 해결하기 위해 기존의 감성사전을 활용한 대용량 학습데이터 자동 구축방안에 대해 제안한다. 감성사전으로는 많이 활용되고 있는 “KOSAC(한국어감성분석코퍼스) 감성사전[1][2]”과 “KNU 한국어 감성사전[2][3]”을 이용해 학습데이터를 구축하였다. 감성사전을 통해 학습데이터를 구축하면 감성이 있는 많은 양의 레이블 된 학습데이터를 추출할 수 있다는 이점이 있다.

제안한 방안의 성능을 평가하기 위해 수작업과 감성사전을 활용했을 경우를 비교하여 실험하였다. 수작업의 경우에는 3명의 평가자가 참여하여 직접 감성을 분류하였고 감성사전의 경우에는 긍정과 부정에 해당하는 감성사전과 문장을 비교해 긍정과 부정의 빈도를 계산해 빈도가 높은 것으로 감성을 분류하였다. 그 다음 딥러닝 기법 중 하나인 앞 뒤 문맥을 살펴볼 수 있는 Bi-LSTM(Bi-directional LSTM)으로 학습한 후 그 모델을 이용한 감성분석을 통해 성능을 평가하였다. 실험 결과 감성사전을 이용한 모델이 더 짧은 시간에 대용량의 학습데이터를 생성할 수 있었고 정확도가 향상되는 것을 볼 수 있었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구

1) http://word.snu.ac.kr/kosac/lexicon.php
2) http://dilab.kunsan.ac.kr/knusl.html

- 이 논문은 2016년도 정부(미래창조과학부)의 한국연구재단의 중견연구자지원사업(No. NRF-2016R1A2B1014843)의 연구비 지원으로 수행하였습니다. 이 논문은 2017년 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2017M3C4A7068188)

* 교신저자 : 온병원

에 대해 기술한다. 3장에서는 감성사전을 활용한 대용량 학습데이터 구축방안에 대해 서술하고 4장에서는 감성분석을 위한 딥러닝 모델에 대해 설명한다. 5장에서는 제안한 모델로 평가한 결과를 분석하고 마지막으로 6장에서는 결론에 대해 기술한다.

2. 관련 연구

한국어 감성분석은 기계학습을 이용한 기법이 많이 연구가 되어져왔다. 기계학습 기반의 감성분석은 감독(Supervised), 비감독(Unsupervised), 반감독(Semi-supervised) 학습에 의해 수행되며 감독학습의 대표적인 기법으로는 나이브 베이즈(Naive Bayes), 지지벡터기계(Support Vector Machine) 등이 있다[4]. [5]에서는 영어 단어 시소러스의 유의어 정보로 단어를 확장하고 이를 번역하여 감정 자질을 추출한 뒤 기계학습 기법 중 하나인 지지벡터기계를 사용하여 문서의 감성을 긍정과 부정에 초점을 맞춰 분류하였다. [6]에서는 트윗에서 자질을 추출한 뒤 지지벡터기계, 나이브 베이즈로 트위터의 감성을 분류해 그 성능을 평가하였다. 최근에는 딥러닝 기법이 점점 발달하면서 딥러닝을 활용한 감성분석도 많이 연구되고 있다. 지지벡터기계, 나이브 베이즈 같은 기존의 기계학습 방식은 학습데이터의 특성에 영향을 많이 받아 도메인 적응에 취약해 저조한 성능을 보이는데 이를 해결하기 위해 딥러닝 기법을 적용하면 학습데이터에서 높은 수준을 특성을 추출하여 뛰어난 성과를 얻을 수 있다[4]. LSTM(Long Short Term Memory) 알고리즘을 사용하여 인스타그램의 비정형 텍스트에 대한 감성분석을 한 모델이 있는데 이 모델에서는 25,000개의 학습데이터를 사용하였다[7].

본 논문에서는 좋은 성과를 얻을 수 있는 딥러닝 기법을 이용해 대용량의 학습데이터를 생성하는 방안에 대해 제안한다. 기존에 구축되어 있는 감성사전의 어휘를 자질로 사용해 충분한 양의 학습데이터를 구축한다. 그리고 그 학습데이터를 딥러닝 기법 중 하나인 Bi-LSTM을 이용해 학습한 후 그 모델을 이용한 감성분석을 통해 제안 모델의 성능을 입증한다.

3. 제안방안 : 대용량 학습데이터 구축방안

수작업, KOSAC 감성사전, KNU 한국어 감성사전을 이용한 학습데이터 구축방안에 대해 설명한다. 본 논문에서는 학습데이터 구축을 위해 네이버 카페 “한국 형태소 분석 등 NLP 연구개발 자료”에서 제공하는 한국어 원시 말뭉치 1억2천3백만 어절(약1천만 문장)을 사용하였다[8]. 이 데이터는 뉴스기사, 수필, 소설 등과 같은 한국어 원시 말뭉치로 이루어져 있다. 형태소 분석기는 꼬꼬마 형태소 분석기[9]로 모두 동일하게 사용하였다.

3.1. 수작업

전체 데이터 중 무작위로 15,000개의 문장을 추출해 그 문장을 긍정과 부정으로 분류하였다. 3명이 분류에 참여하였고 3명의 평가자 중 2명 이상이 동일한 감성으로 판단한 것으로 긍정과 부정을 결정하였다. 표 1은 수작업으로 감성을 분류한 통계이다.

표 1 수작업 분류 통계

감성	개수	비율
긍정	3,785	25.2%
부정	3,620	24.1%

일반적으로 절 하나의 형태소 개수가 30개 이하이므로 긍정과 부정으로 분류된 문장을 형태소 분석을 한 후 형태소 개수가 30개 이하인 문장만 추출하였다. 그 중 긍정과 부정 각 2,200개 씩 4,400개를 추출해 3,800개는 학습에 사용하였고 600개는 평가에 사용하였다.

3.2. 감성사전

KOSAC 감성사전과 KNU 한국어 감성사전으로 대용량 학습데이터를 구축한 과정에 대해 설명한다. 그림 1과 같이 긍정과 부정에 해당하는 사전의 어휘와 문장을 비교해 그 빈도를 계산하고 빈도가 높은 것으로 감성을 분류하였다.

```
Input : 문장 S

if 긍정어휘 수 > 부정어휘 수
        S ← 긍정문장

else 긍정어휘 수 < 부정어휘 수
        S ← 부정문장
```

그림 1 감성분류 기준

```
if 긍정문장
    감성점수 = 긍정어휘 수 - 부정어휘 수

else 부정문장
    감성점수 = 부정어휘 수 - 긍정어휘 수
```

그림 2 감성점수 식

그림 2는 감성점수에 대한 식이다. 긍정어휘 수와 부정어휘 수의 차를 통해 감성점수를 매겼으며 감성을 분류한 문장 중 긍정과 부정 각각 감정점수가 높은 순 대로 각 150,000개 씩 총 300,000개의 학습데이터를 추출하였다. 이 때 긍정과 부정으로 분류된 문장의 형태소 개수가 30개 이하인 문장만 추출하였다.

귀엽다는 생각이 들기 보다는 불쌍하다는 생각이 들어서 저를 우울하게 만들었어요.

그림 3 감성문장 예시

예를 들어 그림 3과 같은 문장이 있을 때 긍정어휘는 “귀엽다”가 있고 부정어휘는 “불쌍하다”, “우울하게”가 있다. 이때 긍정어휘 수는 1, 부정어휘 수는 2로 이 문장은 부정문장으로 판단을 하고 감성점수는 2-1로 1로 매긴다.

3.2.1. KOSAC(한국어감성분석코퍼스) 감성사전

KOSAC(한국어감성분석코퍼스) 감성사전[2]은 서울대학교 언어학과에서 구축한 감성사전으로 1-gram, 2-gram, 3-gram의 어휘로 구성되어 있으며 형태소 단위로 감성어휘를 제공한다.

본 논문에서는 KOSAC 감성사전의 내용 중 극성을 나타내주는 ‘Polarity’ 속성을 사용했으며 그 중에서 ‘max.value’ 값이 긍정을 나타내는 POS와 부정을 나타내는 NEG만을 추출하였다. KOSAC 감성사전은 감성 어휘가 형태소 분석된 상태로 존재한다는 점에서 KNU 한국어 감성사전과 차이점이 있다. 감성 어휘가 형태소 분석된 상태로 존재하기 때문에 문장 전체를 형태소 분석을 한 다음 감성을 분류하였다. 표 2는 KOSAC 감성사전으로 긍정과 부정을 분류한 통계이다.

표 2 KOSAC 감성사전 분류 통계

감성	개수	비율
긍정	4,672,425	44.9%
부정	4,705,665	45.2%

표 3은 감성점수를 기준으로 추출한 학습데이터 문장의 예시이다.

표 3 KOSAC 감성사전 추출 문장 예시

감성	문장	감성점수
긍정	더욱 민화의 선과 색을 되살려낸 삽화들이 시원하고 아름답다.	23
	현실과 환상을 자유롭게 넘나들며 참신한 이야기와 이미지를 만들어낸 독창성도 눈에 띈다.	22
부정	문화의 단절은 골이 점점 깊어지고, 옛글은 자꾸 고리타분하게만 보인다.	23
	잊을만하면 문제가 되는 것을 보면 아직도 잘못된 관행이 사라지지 않았음이 분명하다.	22

3.2.2. KNU 한국어 감성사전

KNU 한국어 감성사전[3]은 군산대학교에서 구축한 한국어 감성사전으로 국립국어원에서 제공하는 ‘표준국어대사전(국립국어원 2018)’을 구성하는 모든 단어와 그 단어에 해당하는 뜻풀이(gloss)를 수집한 후 정제하여 만든 것이다. 그 외에도 축약어, 이모티콘 등 표준국어대사전에서 추출되지 않은 새로운 감성어를 추출하였다. 특정 도메인에서 사용되는 감성어보다는 어떤 도메인에도 사용될 수 있는 보편적인 긍부정어로 구성된다. 긍정, 부정, 중립으로 나누며 1-gram, 2-gram, n-gram, 문형에 해당하는 감성어가 있다.

본 논문에서는 긍정과 부정을 나타내는 어휘만을 추출하였다. 표 4는 KNU 한국어 감성사전으로 긍정과 부정을 분류한 통계이다.

표 4 KNU 한국어 감성사전 분류 통계

감성	개수	비율
긍정	3,279,165	31.5%
부정	2,427,206	23.3%

표 5는 감성점수를 기준으로 추출한 학습데이터 문장의 예시이다. KNU 한국어 감성사전의 어휘들은 형태소가 분석되지 않은 상태로 존재해 감성점수의 범위가 KOSAC 감성사전에 비해 작다.

표 5 KNU 한국어 감성사전 추출 문장 예시

감성	문장	감성점수
긍정	구단은 수익과 함께 긍정적인 팀 이미지를 얻어 좋고 팬들은 색다른 체험을 할 수 있어 반갑다.	9
	정확한 수비 위치 선정을 바탕으로 상대 패스를 차단하는 능력이 뛰어나다는 평가다.	7
부정	불합격이라는 이름으로 그들에게 수치와 죄의식과 불안과 좌절과 패배와 허탈이라는 고통을 안겨주고 있다.	10
	그러니까 이 흉악한 범죄는 의도된 악에서 비롯되지 않아 더 끔찍하다.	9

4. 감성분석을 위한 딥러닝 모델

그림 4는 본 논문에서 제안하는 감성분석 딥러닝 모델

의 구조도이다. 입력으로 문장이 주어지면 수작업, KOSAC 감성사전, KNU 한국어 감성사전으로 문장의 긍정과 부정으로 분류한다. 이 때 수작업은 평가자에 의해서 긍정과 부정을 분류해 소량의 학습데이터를 생성하게 되고 감성사전은 감성점수를 기준으로 대량의 학습데이터를 생성한다. 생성된 학습데이터를 워드 임베딩한 후 딥러닝 기법인 Bi-LSTM으로 학습을 진행한다. 마지막으로 학습된 모델을 기준으로 감성분석을 통해 성능 평가를 수행한다.

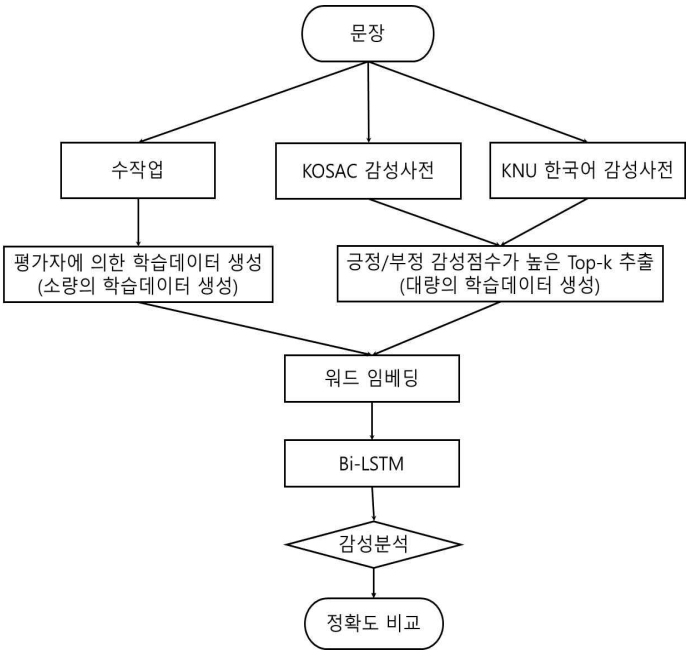

그림 4 감성분석 딥러닝 모델 구조도

4.1. 워드 임베딩

워드 임베딩은 자연어 처리에서 어휘의 단어나 구가 실수의 벡터로 매핑 하는 것으로 딥러닝 모델에 적용할 때 많이 이용된다[10].

Bi-LSTM 모델은 입력으로 벡터 값을 사용하기 때문에 단어를 벡터로 변환해주는 워드 임베딩을 수행하였다. 문장을 형태소 분석한 결과가 워드 임베딩의 입력으로 주어졌으며 이때 형태소 분석기는 꼬꼬마 형태소 분석기[9]를 사용하였다. 형태소 분석한 문장은 형태소와 품사의 결합으로 나타냈으며 형태는 표 6과 같다.

표 6 형태소 분석 결과

좋/VA, 은/ETD, 흐름/NNG

본 논문에서는 Facebook의 fastText[11]로 워드 임베딩을 수행하였다. skipgram을 이용해 차원은 50, window size는 7로 설정하여 임베딩을 수행하였다.

4.2. Bi-LSTM 모델

LSTM은 RNN(Recurrent Neural Network)에서 발생하는 길이가 길어질수록 역전파(Back-propagation) 시 기울기(Gradient) 값이 줄어들어 학습 능력이 떨어지는 것을 보완한 모델이다. Bi-LSTM은 순차적 데이터에서 좋은 성능을 보이며 입력된 데이터에 대해 양방향으로 학습이 가능한 딥러닝 기법이다. 본 논문에서는 수작업과 감성사전으로 분류한 데이터를 학습시키기 위해 양방향으로 입력 정보를 받을 수 있는 Bi-LSTM 모델을 사용하였다.

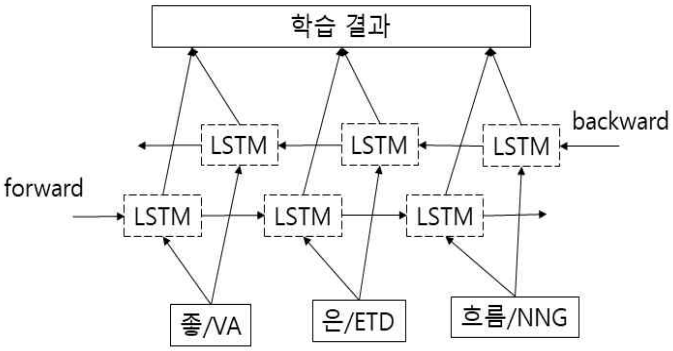

그림 5 Bi-LSTM 예시

그림 5는 본 논문에서 제안한 Bi-LSTM 모델의 예시로 학습 데이터를 워드 임베딩한 결과를 입력으로 학습하고 매번 학습된 결과를 평가데이터에 적용해 정확도 및 결과를 확인하였다.

5. 실험

5.1 실험환경

본 논문에서는 학습을 위한 실험 환경으로 수작업 데이터 모델은 learning rate를 0.00001로 설정하였고 KOSAC 감성사전과 KNU 한국어 감성사전 모델은 learning rate를 0.0001로 설정하였다. batch size는 50, epoch은 12로 동일하게 설정하였다. 그리고 최근 많은 연구에서 좋은 성능을 보이고 있는 adam 최적화 알고리즘[12]을 이용하여 파라미터를 최적화했다. 실험을 위한 모든 모델은 구글에서 오픈소스로 공개한 라이브러리인 Tensorflow[13] 사용하여 구현하였다.

5.2. 성능평가

5.2.1. 학습데이터 구축시간

표 7은 수작업과 감성사전을 이용한 학습데이터 구축에 사용한 총 문장 수를 나타낸 표이다.

표 7 총 문장 수

	총 문장 수
수작업	15,000
KOSAC 감성사전	10,409,444
KNU 한국어 감성사전	10,409,444

표 8은 수작업과 감성사전을 이용한 학습데이터 구축 실행 시간을 나타낸 표이다. 수작업은 총 문장 수인 15,000개에서 3,800개의 학습 데이터를 생성했으며 실행 시간은 3명의 평가자가 걸린 시간으로 나타내었다. KOSAC 감성사전과 KNU 한국어 감성사전은 총 문장 수 인 10,409,444개에서 각 300,000개의 학습데이터를 생성했으며 실행 시간은 전체 데이터에서 감성을 분류한 시간으로 나타냈다. 평균을 낸 결과 수작업, KOSAC 감성사전, KNU 한국어 감성사전 순으로 소요시간이 길다. KOSAC 감성사전은 형태소 분석 된 상태에서 분류하기 때문에 KNU 한국어 감성사전보다 더 많은 시간이 걸렸다. 감성 사전을 이용하면 빠른 시간에 많은 양의 학습데이터 구축이 가능하다는 것을 알 수 있다.

표 8 실행 시간

	소요시간(hh:mm:ss)			평균
수작업	평가자1	평가자2	평가자3	24:42:45
	21:11:35	20:54:02	32:02:40	
KOSAC 감성사전	06:36:36			
KNU 한국어 감성사전	04:20:15			

5.2.2. 정확도

표 9는 평가데이터 600개를 기준으로 정확도를 나타낸 것이다. 수작업의 정확도는 0.5181, KOSAC 감성사전의 정확도는 0.7133으로 수작업으로 했을 경우보다 성능이 약 38% 상승하였고 KNU 한국어 감성사전의 정확도는 0.7567로 수작업으로 했을 경우보다 성능이 약 46% 상승하였다.

수작업의 경우에는 정확한 학습데이터를 구축할 수 있었지만 구축하는데 많은 시간이 소요되어 많은 학습데이터를 생성하는데 한계가 있었다. 또한 적은 양의 학습데이터를 딥러닝 모델 학습에 사용하였기 때문에 정확도가 높지 못한 결과를 보였다. 그에 비해 KOSAC 감성사전과 KNU 한국어 감성사전은 수작업 보다 정확하진 않지만 빠른 시간 내에 많은 양의 학습데이터를 구축할 수 있었으며 이를 통해 딥러닝 모델을 학습한 결과 수작업에 비해 높은 정확도를 보였다. 이와 같이 감성사전을 통해 학습데이터를 구축하는 제안방안이 딥러닝 모델의 성능을 향상시키는데 효과적이라는 것을 알 수 있다.

KNU 한국어 감성사전이 KOSAC 감성사전보다 정확도가 약 0.04 높다. 여러 도메인의 문장이 있는 상태에서 KNU 한국어 감성사전은 도메인에 영향을 받지 않는 사전이기 때문에 정확도가 더 높게 나왔다고 판단한다. 정확도 구하는 식은 아래와 같다.

$$\text{정확도} = \frac{\text{정답데이터 수}}{\text{평가데이터 수}}$$

표 9 정확도

	정확도
수작업	0.5181
KOSAC 감성사전	0.7133
KNU 한국어 감성사전	0.7567

표 10과 표 11은 평가데이터에서 감성을 잘 예측한 각 감성사전의 정답 문장 예시이다.

표 10 KOSAC 감성사전 정답 문장 예시

감성	문장	감성점수
긍정	그 덕에 이정도로 성장할 수 있었다.	4
	코픽스 금리가 인하한 것도 좋은 영향으로 작용하였다.	3
부정	그 때문에 간접적으로 사회 자체의 힘이 약화되었다.	2
	가장 안전해야 할 공간에서 느낀 공포는 몰카 불안증으로 이어졌습니다.	1

표 11 KNU 한국어 감성사전 정답 문장 예시

감성	문장	감성점수
긍정	사장 소리를 듣기에는 아직 젊은 나이지만 그는 업계에서 유능한 경영인으로 손꼽히고 있다.	2
	재미는 물론이고 뜻밖에 마주치는 삶의 풍경 속에서 감동까지 안겨 주는 것이 짧은 소설의 매력이다.	2
부정	상당수 업체들이 경기 침체에 따른 실적 부진과 자금 경색으로 생존마저 위협받았다.	1
	최근 유럽 경제는 전후 최악이라 불릴 만큼 심각하다.	1

6. 결론

본 논문에서는 감성분석에서의 학습데이터 부족 문제 해결을 위해 감성사전을 활용한 대용량 학습데이터 구축 방안을 제안하였다. 감성사전으로는 KOSAC 감성사전과 KNU 한국어 감성사전을 활용했으며 그 성능을 평가하기 위해 수작업과 함께 비교하였다. 수작업은 3명의 평가자가 참여하여 학습데이터를 구축하였고 감성사전은 감성사전에 있는 긍정과 부정 어휘의 빈도가 높은 것으로 감성을 결정해 학습데이터를 구축하였다. 실험결과 감성사전을 활용한 경우 수작업 보다 빠른 시간 내에 많은 양의 학습데이터를 구축할 수 있었다. 또한 감성사전을 활용하여 구축된 학습데이터를 학습시키는 것이 더 높은 정확도를 보였으며 제안방안의 유용성과 실효성을 입증하였다.

향후에는 학습데이터 구축에 구문분석, 의미분석 등의 자질을 추가하여 보다 정교한 학습데이터를 자동으로 구축하는 연구를 수행할 예정이다.

참고문헌

[1] IDG Tech Report, “글에서 감성을 읽다” 감성 분석의 이해, 2014
[2] http://word.snu.ac.kr/kosac/lexicon.php
[3] https://github.com/park1200656/KnuSentiLex (Accessed 2018)
[4] 서상현, 김준태, “딥러닝 기반 감성분석 연구동향”, 한국멀티미디어 학회지, 제20권, 제3호, pp.8-22, 2016
[5] 황재원, 고영중, “효과적인 감성 자질을 이용한 한국어 문서 감정 분류 시스템”, 한국정보과학회 2007 가을 학술발표 논문집, 제34권, 제2호(A), pp.60-61, 2017
[6] 인좌상, 김진만, “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교”, 멀티미디어학회논문지, 제17권, 제2호, pp.232-237, 2014
[7] 손진광, “RNN LSTM과 ACO를 이용한 감성 분석을 통한 컨텐츠 추천 시스템에 관한 연구”, 한국정보과학회 2017 한국소프트웨어종합학술대회 논문집, pp.1033-1035, 2017
[8] https://cafe.naver.com/nlpkang/17
[9] 이동주, 연종흠, 황인법, 이상구, “꼬꼬마 : 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제16권, 제11호, pp.1046-1050, 2010
[10] https://en.wikipedia.org/wiki/Word_embedding
[11] Bojanowski, Piotr, Edouard Grave, and Armand Joulin, Tomas Mikolov, “fastText”, https://research.fb.com/fasttext/
[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014
[13] http://www.tensorflow.org