감정 어휘 사전을 활용한 영화 리뷰 말뭉치 감정 분석

장연지 [O,1], 최지선 [2], 박서윤 [3], 강예지 [3], 강혜린 [3], 김한샘[†]

국립국어원 [1], ㈜이르테크 [2], 연세대학교 [3†]

yeonji3547@korea.kr, wltjs823@iirtech.com, {seoyoon.park, yjkang5009, hyerink, khss}@yonsei.ac.kr

Movie Corpus Emotional Analysis Using Emotion Vocabulary Dictionary

Yeonji Jang[O,1], Jiseon Choi[2], Seoyoon Park[3], Yejee Kang[3], Hyerin Kang[3], Hansaem Kim[†]

National Institute of Korean Language[1], IIRTECH[2], Yonsei University[3†]

요 약

감정 분석은 텍스트 데이터에서 인간이 느끼는 감정을 다양한 감정 유형으로 분류하는 것이다. 그러나 많은 연구에서 감정 분석은 긍정과 부정, 또는 중립의 극성을 분류하는 감성 분석의 개념과 혼용되고 있다. 본 연구에서는 텍스트에서 느껴지는 감정들을 다양한 감정 유형으로 분류한 감정 말뭉치를 구축하였는데, 감정 말뭉치를 구축하기 위해 심리학 모델을 기반으로 분류한 감정 어휘 사전을 사용하였다. 9가지 감정 유형으로 분류된 한국어 감정 어휘 사전을 바탕으로 한국어 영화 리뷰 말뭉치에 9가지 감정 유형의 감정을 태깅하여 감정 분석 말뭉치를 구축하고, KcBert에 학습시켰다. 긍정과 부정으로 분류된 데이터로 사전 학습된 KcBert에 9개의 유형으로 분류된 데이터를 학습시켜 기존 모델과 성능 비교를 한 결과, KcBert는 다중 분류 모델에서도 우수한 성능을 보였다.

주제어: 감정 분석, 감정 어휘 사전, NSMC, KcBert

1. 서론

감정 분석(Emotional Analysis)은 텍스트에서 느껴지는 감정을 다양한 감정 유형으로 분류하는 것이다. 많은 연구에서 감정 분석의 개념은 감성 분석(Sentiment Analysis) 또는 오피니언 마이닝(Opinion mining)과 혼용되어 쓰이고 있는데, 엄밀히 말하자면 감성 분석과 감정 분석은 변별되는 개념이다.

극성(Polarity)을 판단하여 긍정과 부정 또는 중립으로 분류하는 감성 분석과는 달리, 감정 분석은 어떤 대상이나 상황, 분위기 등에 대해 갖게 되는 기쁨, 슬픔, 분노 등의 느낌을 판단하여 여러 감정 유형으로 분류하는 것이다. 따라서 개인의 의견을 좋고 싫음으로 표현하는 오피니언 마이닝이나 감성 분석과는 구분되는 개념으로 볼 수 있다. 이러한 연구에서는 SNS 게시글이나, 일기, 대화, 리뷰 등 인간의 감정이 다양하게 잘 드러나는 데이터를 확보해야 한다. 사람들이 직접 영화를 보고 작성한 리뷰를 수집하여 긍정과 부정, 중립으로 분석을 한 NSMC(Naver Sentiment Movie Corpus)[1] 역시 사람의 감정이 잘 드러나는 데이터를 기반으로 구축한 감성 분석 말뭉치이다. 그러나 영화 리뷰 데이터의 경우 특정 영화에 대한 좋고 싫음을 표현하는 것 이상으로 영화를 보고 느낀 본인의 감정을 구체적으로 표현한 것이기 때문에 긍부정 이상의 구체적인 감정 분석이 필요하다. 이에 본 연구에서는 뉴스나 백과 사전 등에 비해 보다 다양한 감정이 표현되어 있는 영화 리뷰 데이터를 대상으로 감정 분석 말뭉치를 구축하였다. 감정 분석을 진행하는 과정에서 대규모의 영화 리뷰 데이터에 다양한 감정 유형을 태깅하기 위해 감정 어휘 사전을 활용하였다. 또한, 언어 모델 중 긍·부정으로 이진 분류된 댓글 데이터를 기반으로 학습한 KcBert(Korean comments Bert)에 9가지 감정 유형으로 태깅된 감정 분석 말뭉치를 학습시켜 성능을 평가하였다.

2. 관련 연구

2.1. 감성 및 감정 어휘 사전

최근 딥러닝 모델을 활용한 언어 연구에서 모델의 정확도 향상을 위해 대규모 학습 데이터를 구축하는데, 학습 데이터 구축에 어휘 사전이 다양하게 활용되고 있다. 본 연구에서도 감정 분석 말뭉치를 구축하기 위해 감정 어휘 사전을 활용하였는데, 양질의 감정 분석 데이터를 구축하기 위해서는 데이터에 가장 적합한 어휘 사전을 찾는 것이 관건이다. 감정 분석에 활용하기 위해 기구축된 다양한 어휘 사전 중에서도 감정 어휘 사전보다 감성

[1] https://github.com/e9t/nsmc

어휘 사전을 더 많이 찾아볼 수 있었다.

먼저, 대표적인 영어 감성 사전으로 SentiWordNet은 프린스턴 대학에서 영어 의미 어휘 목록 구축 프로젝트인 WordNet을 기반으로 한 것으로, synset이라는 유의어 집단에 포함된 단어들을 유의어, 반의어 관계를 통해 확장하고 이를 분류기로 학습하여 긍정, 부정, 객관성에 대한 값을 부여한 대표적인 감성사전이다.[2] 그러나 SentiWordNet은 한국어로 번역하여 사용할 때 한국어에서는 다양한 표현으로 나타나는 어휘들이 영어로 같은 어휘로 대역되는 문제가 있다[3]. SentiWordNet을 이용하여 구축된 DecoSelex는 한국어 Deco 사전을 이용해 확장하는 방식으로 오피니언 마이닝을 위해 구축한 감정 사전이나, 공개되어 있는 사전은 아니다[4]. 오픈 서비스로 제공되었던 감정 사전인 '오픈 한글'은 집단지성을 이용한 감성어휘 사전으로, 다수의 참여자가 특정 단어에 대해 긍정, 부정, 중립으로 투표한 결과를 기반으로 구축되었는데, 현재 잠정적으로 서비스가 중단된 상태이다[5]. 이외에도 서울대에서 구축한 KOSAC 말뭉치를 기반으로 한 한국어 감성 어휘 목록(감성사전)[2], 한국외국어대학교에서 오피니언 마이닝을 위해 구축한 MUSE(Multilingual Sentiment Lex ica & Sentiment-Annotated Corpora) 감성 사전 SELEX[3] 등이 있다.

지금까지 살펴본 바와 같이 특정 어휘에 대한 사람의 판단이나 어휘 정보를 통해 극성을 판단하여 감성 어휘 사전을 구축하는 연구 외에 언어 모델을 기반으로 한 감성 사전 구축 연구도 이루어졌다. [3]에서는 국립국어원 표준국어대사전의 단어 뜻풀이에 대해 3명의 투표자가 감성 평가를 하여 학습용 데이터를 구축한 후 Bi-LSTM 모델을 학습하였는데, 모델 학습을 통해 긍정, 부정 어휘를 분류하고 단어를 n-gram으로 추출하여 KNU 한국어 감성 사전을 구축하였다. 또한 감성 단어사전 및 신조어, 이모티콘 등을 추가하였다. [6]에서는 BERT 토큰화 모델을 사용하여 감정 사전을 구축하였는데, 감정 사전 구축을 위해 서강대학교 감정 텍스트 데이터 셋을 이용하여 '행복', '중립', '슬픔', '분노' 네 가지 감정 문장을 나누고, 각 문장의 감정 유형에 따라 문장에 나타나는 모든 토큰에 대해서 해당 감정에 영향을 미치는 정도에 대해 0부터 1사이의 값을 주었다.

한편, 심리 모델이나 감정 분류를 중심으로 하는 감정 어휘 사전 연구도 이루어졌는데[9-11], 감정동사에 대해 통사적, 의미적 특성을 바탕으로 '동정, 수치심, 기쁨, 노여움, 슬픔, 두려움, 좋아함, 싫어함, 바람'의 의미 영역으로 나누는 연구가 있었으며[7], 한국어 텍스트 데이터가 긍·부정의 레이블링 수준을 크게 벗어나지 못하고 있어 "AI Hub 한국어 감정 정보가 포함된 단발성 대화 데이터셋"[4]에서는 다분류 감정에 대한 텍스트 데이터셋 필요성에 따라 공포, 분노, 슬픔, 혐오, 행복, 놀람, 중립의 7개 감정 분류를 레이블링한 38,594문장을 제공하고 있다. [8]에서는 러셀의 핵심 정서(Russell core affect)심리 모델을 적용하여 다중 감정을 결정하는 모델을 제안하기도 하였는데, 해당 모델에 "AI Hub 한국어 감정 정보가 포함된 단발성 대화 데이터셋"을 학습용 데이터로 사용하였다. [9]은 연세대학교에서 제작한 '현대 한국어의 어휘빈도' 자료집으로부터 빈번하게 사용하는 감정단어들을 추출하고 감정연구자 12명이 감정표현 어휘의 범주와 표현강도를 파악하는 작업을 수행하였다. Ekman[1]이 정의한 기본 감정 여섯 가지와 HCI에서 활용도가 높은 세가지 감정 범주를 기준으로 하였다. 그 결과 총 504개의 감정표현단어들을 '기쁨', '슬픔', '공포', '분노', '혐오', '놀람', '흥미', '지루함', '통증', 중성', '기타' 범주로 나누어 제시하였다. 이 외에도 경희대학교에서 '한국어 감정표현단어의 추출과 범주화'의 프로젝트 수행을 위해 BK21 플러스에 업로드 된 감성 단어 목록[5]이 있는데, 해당 감성 단어 목록은 총 428개의 단어로 구성되어 있으며 감성 범주, 빈도, 감성 정도로 구성되어 있다.

본 연구에서는 [9]에서 제시한 감정 어휘 사전을 활용하여 감정 분석 말뭉치를 구축하고, 데이터를 언어 모델에 적용하여 다양한 유형의 감정이 태깅된 데이터의 학습 성능을 확인하였다.

2.2. 감정 분석을 위한 언어 모델

자연어처리 분야에서도 감정 분석 연구 분야가 각광을 받고 있으나 한국어와 관련해서는 대부분 극성을 분류하는 감성 분석이 주를 이루고 있어 한국어 감정 분석 연구를 찾아보기 어렵다. 지금까지 자연어처리 분야에서 RNN, CNN, LSTM, ELMo(Embedding from Language Model), BERT 등 다양한 모델을 제안하는 연구들이 이루어졌는데, 각 모델들은 서로 다른 말뭉치를 학습한 것으로 대부분 문어체 텍스트로 학습되었다. Bert 중에서도 대표적으로 Google의 BERT-multilingual과 SKT의 KoBERT가 문어 텍스트를 학습한 언어 모델이라고 할 수 있다. BERT-multilingual은 한국어가 포함된 100개의 언어로 작성된 위키피디아 문서를 학습한 모델이며, KoBert는 기존 한국어 위키피디아 문서를 학습한 ERT-multilingual 모델에 위키피디아 500만 문장, 그리고 한국어 뉴스 데이터 2000만 문장을 다시 학습시킨 모델이다. Bert 외에도 TransFormer를 사용하여 사전 학습을 한 KoELECTRA, SKT AI의 KoGPT2 등도 있다.

지금까지 공개된 Bert 모델들은 대부분 문어 데이터를 학습하여 특히 문어체에 특화된 결과를 보여준다. 따라

[2] http://word.snu.ac.kr/kosac/index.php

[3] http://dicora.hufs.ac.kr

[4] https://www.aihub.or.kr/keti_data_board/language_intelligence

[5] http://datascience.khu.ac.kr/board/bbs/board.php?bo_table=05_01&wr_id=91

서 일반 사람들이 작성하는 댓글이나 채팅 등의 비정형 데이터에도 대응할 수 있도록 한국어 댓글 데이터를 기반으로 학습한 KcBert와 Tokenizer가 개발되었다. KcBert는 2019년 1월부터 2020년 6월까지의 네이버 뉴스 댓글 약 1억 1천만 건을 수집하여 최소한의 전처리 과정을 거쳐 Huggingface의 Bert WordPiece 토크나이저를 학습한 모델이다. KcBert는 Base와 Large 모델의 학습을 진행하여 Huggingface의 Transformers 오픈소스로 공개되어 있다[13].

본 연구에서는 한국어 영화 리뷰 말뭉치를 사용하여 리뷰에 드러나는 감정을 분석하기 때문에 댓글로 학습되어 비정형 데이터에 적합한 결과를 보이는 KcBert 모델을 사용하였다.

3. 감정 어휘 사전을 사용한 감정 분석 데이터

3.1. 감정 분석 데이터

NSMC(Naver Sentiment Movie Corpus v1.0)는 한국어로 된 영화 리뷰 말뭉치로써 포털 사이트 '네이버'의 '네티즌 평점'을 바탕으로 구축된 말뭉치이다. NCMC는 '네티즌 평점' 리뷰의 원 Id와 리뷰 본문 내용(document), 그리고 리뷰 내용에 대한 레이블(label)로 구성되었다. 말뭉치 규모는 총 20만 건이며, https://github.com/e9t/nsmc/에서는 train 15만 건, test 5만 건으로 구분하여 배포한다.

NSMC는 id, document, label로 구성되어 있는데, id는 네이버가 제공하는 리뷰 아이디, document는 실제 사람들이 작성한 영화 리뷰, label은 리뷰를 긍정과 부정으로 나눈 극성을 나타낸다. 긍정 레이블 수와 부정 레이블 수는 각 10만 개로 동일한 분포를 나타낸다. 긍정 레이블의 경우 '네티즌 평점'의 원 리뷰가 9~10점일 경우, 부정 레이블의 경우 1~4점일 경우 부착되었다. 두 점수대에 속하지 않은 5~8점 대는 중립(neutral)으로 간주하여 NSMC에서는 제외되었다. 본 연구는 긍정 혹은 부정의 이진분류를 채택한 감성분석 기반의 NSMC의 레이블 대신 감정 어휘사전을 기반으로 하여 9가지 유형의 레이블을 부착하여 감정분석을 진행하였다.

감정 분석 데이터 구축에 사용된 감정 어휘 사전은 [9]을 인용하여 구축된 총 428개의 감정 어휘 목록으로, 어휘들은 총 11가지의 감정 유형(분노, 공포, 흥미, 놀람, 기쁨, 슬픔, 지루함, 통증, 혐오, 중성, 기타)으로 분류되어 있다. [9]에서 제시한 감정 유형의 범주는 [1]가 정의한 기본 감정 6가지 감정(기쁨, 슬픔, 공포, 분노, 혐오, 놀람)에 HCI에서 활용도가 높은 3가지 감정(흥미, 지루함, 통증)을 더하였다.

본 연구에서는 감정 어휘 사전에 있는 어휘 목록을 형태 분석하여 네이버 영화 리뷰 데이터와 일치하는 것에 해당 감정 유형을 부착하는 방식으로 9가지 감정 유형의 감정 분석 데이터를 구축하였다.

3.2. 감정 어휘 사전 기반 감정 태깅

앞서 서술한 바와 같이 본 연구는 감정 어휘 사전을 활용하여 감성 분석 기반 말뭉치를 감정 분석 기반 말뭉치로 가공하였다. 감정 분석 데이터 구축에 사용된 감정 어휘 사전은 [9]을 인용하여 구축된 총 428개의 감정 어휘 목록으로, 어휘들은 총 11가지의 감정 유형(분노, 공포, 흥미, 놀람, 기쁨, 슬픔, 지루함, 통증, 혐오, 중성, 기타)으로 분류되어 있다.

먼저, 감정 태깅을 위해 감정 어휘 사전의 어휘 목록을 형태 분석하여 형태 정보를 추가하였다. 형태 분석된 감정 어휘 사전의 구조는 표 1과 같다.

표 1. 형태 분석된 감정 어휘 사전 구조

Id	Word	Em	St	Pos
3	가뿐하다	기쁨	1	[('가뿐', 'XR'), ('하', 'XSA'), ('다', 'EFN')]
95	낙담하다	슬픔	0	[('낙담', 'NNG'), ('하', 'XSV'), ('다', 'EFN')]
326	잔인하다	공포	0	[('잔인', 'NNG'), ('하', 'XSA'), ('다', 'EFN')]
164	분개하다	분노	0	[('분개', 'NNG'), ('하', 'XSV'), ('다', 'EFN')]
416	흉측하다	혐오	0	[('흉측', 'NNG'), ('하', 'XSA'), ('다', 'EFN')]
39	경악하다	놀람	0	[('경악', 'NNG'), ('하', 'XSV'), ('다', 'EFN')]
59	궁금하다	흥미	1	[('궁금', 'XR'), ('하', 'XSA'), ('다', 'EFN')]
23	갑갑하다	지루함	0	[('갑갑', 'XR'), ('하', 'XSA'), ('다', 'EFN')]
46	고통스럽다	통증	0	[('고통', 'NNG'), ('스럽', 'XSA'), ('다', 'EFN')]

표 1과 같이 형태 정보가 추가된 어휘 사전을 활용하여 감정 분석 말뭉치를 구축하였는데, 감정 태깅은 형태 분석 결과를 기준으로 네이버 영화 데이터에 포함된 감정 어휘를 추출하고 각 감정 어휘의 감정 유형을 태깅을 하는 방식으로 진행하였다.

또한, 기존 네이버 영화 말뭉치의 이진 분류를 참고하여 감성 분석 결과가 긍정인 데이터에는 '기쁨', '흥미'에 해당하는 어휘 사전을 연결하였고, 감성 분석 결과가 부정인 데이터에는 '슬픔', '공포', '분노', '혐오', '놀람', '지루함', '통증'에 해당하는 어휘를 연결하였다. 감성 분석의 극성이 감정 유형과 온

전히 일치하는 것은 아니지만, 부정적인 의견을 제시한 리뷰에 단순 감정 어휘 일치로 인해 잘못된 감정이 연결되지 않도록 방지하고자 하였다. 따라서 극성값을 참고하여 기본적으로 긍정적인 감정 유형과 부정적인 감정 유형을 나눈 이후에 일부 일치하지 않는 어휘들은 수정 작업을 거쳐 감정 태깅의 일관성을 유지하도록 하였다.

아래 표 2는 본 연구에서 태깅한 감정 유형별 레이블 빈도이다. 감정 어휘 사전을 기반으로 감정 유형을 태깅한 영화 리뷰 데이터는 총 77,273건이며, 감정 유형별 분포에는 다소 차이가 있다. 다음으로 구축한 데이터를 언어 모델에 학습시키기 위해 감정 유형 레이블을 0부터 8까지 숫자로 치환하였다.

표 2. 감정 분석 말뭉치의 감정 분포

감정 레이블	데이터 수 (건)
기쁨(0)	42,328
슬픔(1)	8,070
공포(2)	1,280
분노(3)	11,361
혐오(4)	2,443
놀람(5)	3,325
흥미(6)	480
지루함(7)	7,831
통증(8)	155
전체	**77,375**

아래 표 3은 감정 어휘 사전을 기반으로 감정 유형을 태깅한 네이버 영화 데이터의 예시이다.

표 3. 감정 태깅을 한 네이버 영화 데이터

Id	Document	Label
5805507	촬영영상들이 정말 훌륭합니다만... 불만이 있다면 더빙이 정말 **역겹네요**.	4
10248418	애니는 작화도 좋고 색감도 **좋고 재미있는데**,	1
2989021	시도는 좋았지만.. 진짜 **재미없네**..	7
9684232	중반이후 도무지 공감할 수 없는 스토리 전개에 지루하다 못해 **짜증**이 난다.	3

4. 실험 및 결과

본 연구에서는 감정 어휘 사전을 기반으로 네이버 영화 리뷰 데이터에 감정을 태깅한 후 여러 유형으로 태깅된 감정 분석 말뭉치를 언어 모델에 학습시켜 멀티 레이블 데이터의 학습 성능을 평가해 보았다. 감성 분석 말뭉치는 총 77,273건이며 훈련 데이터는 61,818건, 테스트 데이터는 15,455건이다.

4.1. KcBert 모델 학습

KcBERT(Korean comments BERT)는 NSMC와 같은 정제되지 않은 구어 데이터에 적용하기 위해 15GB의 뉴스 댓글 데이터를 수집하여 Tokenizer와 Bert 모델을 사전 학습한 언어 모델이다[13].

감정 분석 데이터의 학습 성능 평가를 위해 구축한 훈련 데이터와 테스트 데이터를 HuggingFace의 Tokenizers 라이브러리를 통해 Bert WordPiece Tokenizer를 학습시켰다. 모델을 생성할 때 KcBERT에 기본값으로 설정된 입력과 출력의 Label 수가 2개로 설정되어 있어 멀티 레이블 분류를 위해 해당 값을 태깅된 감정 유형의 수에 따라 9로 수정하였다.

본 연구에서는 KcBert-Base 모델과 KcBert-Large 모델에 감정 분석 데이터를 학습시켰으며, 학습한 두 모델의 파라미터는 다음과 같다.

표 4. KcBert 실험 하이퍼파라미터

Model	KcBert-Base	KcBert-Large
Batch size	32	32
Learning rate	2e-5	5e-5
Validation Batch size	32	32
Warm-up	10	40,000

4.2. 학습 결과

표 5는 본 연구에서 구축한 감정 분석 말뭉치를 학습시킨 KcBert의 base 모델과 Large 모델의 성능 비교이다.

표 5. 데이터별 Accuracy

모델	데이터	Accuracy
KcBERT-Base	[13]의 NSMC 이진 분류	89.62
	본 연구의 다중 분류	**91.63**
KcBERT-Large	[13]의 NSMC 이진 분류	90.68
	본 연구의 다중 분류	**90.92**

본 연구에서는 [13]의 NSMC 이진 분류 데이터로 사전 학습했던 KcBert 모델에 감정 유형을 9가지로 분류하여 구축한 다중 분류 감정 분석 데이터를 학습시킨 것이므로 기존 모델의 성능과도 함께 비교하였다. 비교 결과 KcBert-Base 모델과 KcBert-Large에서 모두 [13]의 이진 분류 데이터보다 본 연구의 다중 분류 데이터에서 성능이 우수함을 알 수 있었다. 다만, 다중 분류 데이터를 학습시켰을 경우 Base 모델이 Large 보다 성능이 우수했다. 이는 Large 모델이 Base 모델에 비해 더 복잡한 구조로 되어 있기 때문에 학습을 시킬 때 더 많은 데이터로 학습을 시켜야 하는데, Base 모델과 같은 양의 데이터를 학습시켰기 때문인 것으로 보인다.

표 6는 본 연구에서 구축한 감정 어휘 사전을 활용한

감정 분석 말뭉치를 학습시킨 KcBert 모델의 F1 Score이다.

표 6. KcBert F1 Score

	KcBert-Base	KcBert-Large
Macro F1	**85.96**	84.22
Weighted F1	**91.14**	91

본 연구에서는 다중 분류의 성능을 평가하기 때문에 감정 범주의 레이블만큼 F1 Score가 산출된다. 따라서 기존 binary 분류기에서 사용하는 성능 평가 지표 대신 모든 범주의 평가값의 평균 점수를 산출하는 Macro Average와 각 감정 레이블마다 데이터 수의 차이가 있다는 점을 감안하여 Weighted Average를 사용하였다. 평가 결과 F1 Score 역시 미세하지만 Base 모델이 Large 보다 성능이 우수했다. F1 Score 역시 Large 모델에 Base 모델과 같은 양의 데이터를 학습시켰기 때문인 것으로 보인다. 본 연구에서 진행한 학습 데이터의 규모가 적기 때문에 성능이 Base 모델이 적합하나, 데이터의 규모가 더 커지면 Large 모델이 적합할 것이다.

5. 결론

본 연구에서는 감정 어휘 사전에서 분류한 9가지의 감정 유형과 감정 어휘를 기반으로 감정 어휘 말뭉치를 구축하였다. 그리고 긍정과 부정으로 분류된 영화 리뷰 데이터로 학습한 KcBert(Korean comments Bert)에 9가지 감정 유형으로 태깅된 감정 분석 말뭉치를 학습시켜 성능을 평가하였다. 평가 결과 이진 분류 데이터로 사전학습을 한 기존의 KcBert 모델만큼 다중 분류 데이터에서도 우수한 성능을 보였으며, KcBert의 Base 모델이 Large 모델보다 더 나은 성능을 보였다. 본 연구에서 사용한 데이터의 규모가 작기 때문에 Base 모델에 더 적합하며, 향후 데이터 규모를 늘려 학습을 시킨다면 Large 모델에서 더 좋은 성능을 보일 것이다.

최근 자연언어처리 분야에서 감성 분석 연구와 관련해 다양한 어휘 사전이 구축되었으나, 본 연구에서는 극성 분류인 감성 분석이 아니라 인간이 느끼는 세부적인 감정을 분류하는 것을 목표로 하여 심리학 기반으로 구축된 감정 어휘 사전을 활용하였다. 그러나 감정 어휘 사전의 어휘가 한 어절로 구성되었다는 점과 구축된 어휘의 양이 많지 않다는 점에서 향후 다양한 감정 태깅을 위한 감정 어휘 사전 구축의 필요성을 확인하였다.

감정 유형 태깅을 감정 어휘 사전의 어휘를 형태 분석하여 댓글 데이터에 있는 어휘와 일치하는지 여부를 바탕으로 수행하였으므로 감정 태깅 대상인데도 불구하고 누락되는 데이터가 있었고, 관용 표현이나 반어법을 인식하지 못하는 경우가 있었다. 이런 경우 본고에서는 태깅 후처리를 통해 보완하였는데 향후에는 데이터에 나타나는 관용표현이나 n-gram 구성의 어휘도 고려하여 감정 분석을 위한 어휘 사전이 보완되어야 할 것이다. 이를 위해 기구축된 어휘 사전과 감정 유형을 결합하는 방법을 고안하는 것이 감정 분석 언어 자원 구축과 관련한 중요한 향후 과제이다.

참고문헌

[1] Ekman, P., “Universals and cultural differences in facial expressions of emotion”, Proceedings of the 1971 Nebraska Symposium on Motivation, 207-283, 1971.

[2] Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, “SentiWordNet 3.0: An Enhanced Lexical R esource f or Sentiment Analysis and Opinion Mining”

[3] 박상민, 나철원, 최민성, 이다희, 온병원, “Bi-LSTM 기반의 한국어 감성사전 구축 방안.”, 지능정보연구, vol 24. No 4, pp. 219-240, 2018.

[4] 신동혁, 조동희, 남지순, “ 한국어 감성 사전 DecoSelex 구축을 위한 영어 SentiWordNet 활용 및 보완 논의”, 한국사전학, No.28, pp.75-111. 2016.

[5] 안정국, 김희웅, “한글 감성어 사전 API 구축 및 자연어 처리의 활용”, 한국지능정보시스템학회 학술대회논문집, pp.177-182. 2014.

[6] 이가은, “Bert 기반 한국어 감정 사전을 이용한 감정 예측기 개발”, 석사, 서강대학교 정보통신대학원, 2020.

[7] 김은영, “국어 감정동사 연구”, 박사, 전남대학교 대학원, 2004

[8] 김해준, 도준호, 선주오, 정서희, 이현아, “감정 분석에서의 심리 모델 적용 비교 연구”, 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp.450-452. 2020.

[9] 손선주, 박미숙, 박지은, 손진훈, “한국어 감정표현 단어의 추출과 범주화”, 감성과학, Vol. 15, No. 1, pp. 105-120, 2012.

[10] 김홍진, 김담린, 김보은, 오신혁, 김학수, “감성 단어 등장 순서를 고려한 영화 리뷰 감성 분석”, 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp.313-316. 2020.

[11] 이준범“KcBert: 한국어 댓글로 학습한 Bert”, 제32회 한글 및 한국어 정보처리 학술대회 논문집, pp.437-440. 2020.