레터논문 (Letter Paper)
방송공학회논문지 제27권 제6호, 2022년 11월 (JBE Vol.27, No.6, November 2022)
https://doi.org/10.5909/JBE.2022.27.6.940
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

라디오 청취자 문자 사연을 활용한 한국어 다중 감정 분석용 데이터셋 연구

이 재 아[a)], 박 구 만[a)‡]

A Study on the Dataset of the Korean Multi-class Emotion Analysis in Radio Listeners' Messages

Jaeah Lee[a)] and Gooman Park[a)‡]

요 약

본 연구에서는 직접 수집한 라디오 청취자 문자 사연을 활용하여 한국어 문장 감정 분석을 수행하기 위한 한국어 데이터셋을 구성하였으며 그 특성을 분석하였다. 딥러닝 언어모델 연구가 활발해지면서 한국어 문장 감정 분석에 관한 연구도 다양하게 진행되고 있다. 그러나 한국어의 언어학적 특성으로 인해 감정 분석은 높은 정확도를 기대하기 어렵다. 또한, 긍정/부정으로만 분류되도록 하는 이진 감성 분석은 많은 연구가 이루어졌으나, 3개 이상의 감정으로 분류되는 다중 감정 분석은 더 많은 연구가 필요하다. 이에 대해 딥러닝 기반의 한국어에 대한 다중 감정 분석 모델의 정확도를 높이기 위한 한국어 데이터셋 구성에 관한 고찰과 분석이 필요하다. 본 논문에서는 설문조사와 실험을 통해 감정 분석이 실행되는 과정에서 한국어 감정 분석이 어떤 이유 때문에 어려운지 분석하고 정확도를 향상시킬 수 있는 데이터셋 조성에 대한 방안을 제시하였으며 한국어 문장 감정 분석에 근거로 활용할 수 있게 하였다.

Abstract

This study aims to analyze the Korean dataset by performing Korean sentence Emotion Analysis in the radio listeners' text messages collected personally. Currently, in Korea, research on the Emotion Analysis of Korean sentences is variously continuing. However, it is difficult to expect high accuracy of Emotion Analysis due to the linguistic characteristics of Korean. In addition, a lot of research has been done on Binary Sentiment Analysis that allows positive/negative classification only, but Multi-class Emotion Analysis that is classified into three or more emotions requires more research. In this regard, it is necessary to consider and analyze the Korean dataset to increase the accuracy of Multi-class Emotion Analysis for Korean. In this paper, we analyzed why Korean Emotion Analysis is difficult in the process of conducting Emotion Analysis through surveys and experiments, proposed a method for creating a dataset that can improve accuracy and can be used as a basis for Emotion Analysis of Korean sentences.

Keyword : Emotion Analysis, Natural Language Processing, Dataset, Multi-class, Korean language model

a) 서울과학기술대학교 나노IT디자인융합대학원(Graduate School of Nano IT Design Fusion, Seoul National University of Science and Technology)
‡ Corresponding Author : 박구만(Gooman Park)
E-mail: gmpark@seoultech.ac.kr
Tel: 82-2-970-6430
ORCID: https://orcid.org/0000-0002-7055-5568
· Manuscript August 31, 2022; Revised September 22, 2022; Accepted September 22, 2022.

Ⅰ. 서 론

딥러닝의 발전으로 감정 분석에 관한 다양한 연구가 진행되고 있다.[1] 한국어는 모호한 표현이 많고, 문맥에 따라 의미가 달라 인공지능이 분석하기 어려운 언어이다. 이에 한국어 감정 분석의 정확도를 높이기 위해 한국어 감정 분석이 어려운 이유에 대해 고찰해보고, 실제 환경에서 수집된 한국어 데이터셋의 특징을 분석할 필요가 있다. 이를 바탕으로 정확한 라벨링이 적용된 데이터셋을 구성해야 한다.

이에 본 연구에서는 한국어 데이터셋을 연구하기 위해 설문조사와 실험을 진행하였다. 첫째로, 감정을 판단하기 어렵다고 생각되는 라디오 청취자 문자 사연을 추려 사람들은 어떤 감정으로 해석하는지 설문조사를 진행하였다. 응답자가 가장 많이 선택한 감정을 보편적인 감정으로 정의하여 데이터셋에 대한 감정을 라벨링하였다. 둘째로, 두 종류의 데이터셋을 감정 분석 모델에 각각 학습하여 실험하였다. 설문 결과를 기준으로 감정이 라벨링된 라디오 청취자 문자 사연을 모델에 학습한 뒤 감정 분석을 수행한 결과와 개방 데이터셋을 학습한 뒤 감정 분석을 수행한 결과를 비교 분석하였다.

본 연구는 실제 환경에서 수집한 데이터로 감정 분석을 수행하였으며, 한국어 데이터셋의 모호성을 줄이기 위하여 설문조사로 감정 라벨링의 기준을 결정하고 혼동 가능성을 줄였다. 본론에서는 서론에서 설명한 설문조사와 실험에 관하여 자세히 서술하고 결론에서는 설문조사와 실험 결과 분석을 통해 추후 연구 방향을 제시한다.

Ⅱ. 본 론

실험하기에 앞서, 그림 1을 통해 본 연구의 실험 시스템에 대하여 간단히 설명하고자 한다.

데이터셋 전처리 후, 훈련 데이터와 테스트 데이터로 나누어 훈련 데이터로 모델을 학습하고, 테스트 데이터로 모델을 평가한다. 이렇게 학습된 모델에 감정을 분석하고자 하는 새로운 문장 데이터를 주입하여 결과를 분석하는 시스템이다. KoBERT[2]는 감정 분석에 뛰어난 성능을 보이며[3,4], 문어체에 특화되어 라디오 청취자 문자 사연 감정 분석을 수행하기에 적합하다.

본 연구에서는 한국어의 언어학적 특성을 분석하고, 감정 분석이 어려운 이유를 고찰한다. 또한, 한국어 감정 분석이 어려운 이유에 대해 고찰하는 데에 그치지 않고 정확도를 높이기 위해 실험을 진행하였다. 실험에 앞서, 인공지능이 인간 감정을 어떻게 감정을 분류해야 하는지에 대한 명확한 기준을 제시하기 위해 인간이 느끼는 보편적인 감정을 파악하고자 설문조사를 진행하였다. 설문 결과를 기반으로 데이터셋을 만들고[5], 감정 분석 모델을 학습하여 감정 분석 테스트를 진행하였다. 설문조사를 기반으로 감정을 라벨링한 후 감정 분석을 수행한 결과와 개방 데이터를 활용한 감정 분석 결과를 비교하여 데이터셋의 특성이 감정 분석 모델의 성능에 끼치는 영향을 분석할 것이다.

1. 개방 데이터셋/라디오 청취자 문자 사연 데이터셋 비교 실험

1.1 감정 분석이 어려운 한국어의 언어학적 특성

한국어 감정 분석이 다른 언어에 비해 어렵다는 많은 연구 결과가 존재한다.[6] 한국어는 같은 단어라도 문맥상 해석이 달라질 수 있어 알고리즘으로 구현되기 어려운 언어이다. 한국어 감정 분석이 어려운 이유는 크게 두 가지로 나뉜다. 첫째로는 상황에 따라 문장의 감정이 달라질 수 있는 경우고, 둘째로는 모호한 감정 표현들이 사용된 경우다.

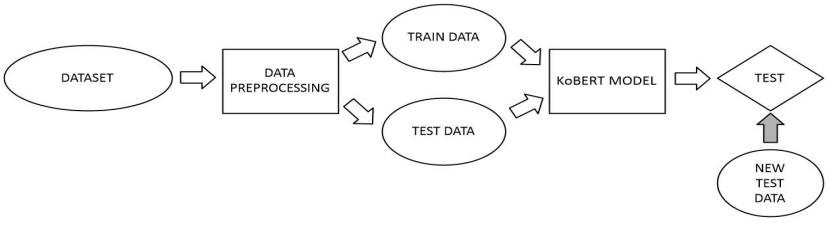

그림 1. 기본적인 실험 시스템 구성도
Fig. 1. Fundamental block diagram of experiment

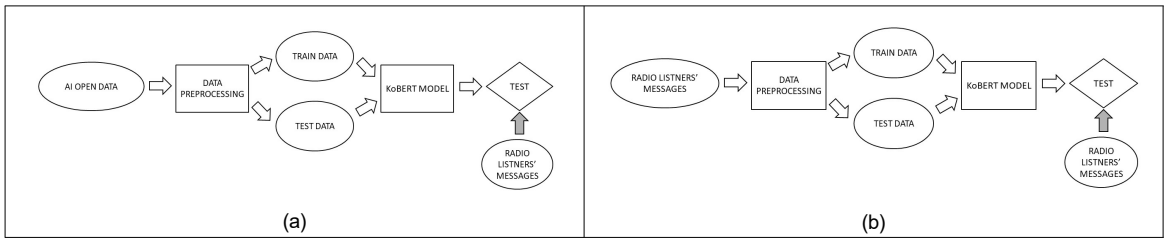

그림 2. 실험 시스템 구성도. (a) : 개방 데이터셋 시스템 구성도, (b) : 라디오 청취자 사연 시스템 구성도
Fig. 2. Block diagram of experiment. (a) : Block diagram of experiment using Open Data, (b) : Block diagram of experiment using Radio Listeners's message

그 예시는 표 1과 같다.

표 1. 감정 분석이 어려운 한국어 문장의 예시
Table 1. Examples of Difficult Korean sentences in Emotion Analysis

case 1 : situational emotion	'만사가 귀찮아요' '참 재미있는 일이 없어서 큰일입니다' '퇴근했어요. 배고파서 어지러워요'
case 2 : ambiguous expression	'정년퇴직을 하게 되니 시원섭섭합니다' '내일 딸 결혼식이라 마음이 싱숭생숭합니다.' '상견례를 하는데 마음이 뭉클했어요'

감정 분석의 경우에는 클래스가 많아질수록 감정 분석의 정확도는 낮아질 수밖에 없다. 또한, 감정을 세부적으로 분류하기 위해 클래스를 늘린다고 하더라도 한국어 감정 분석 클래스의 기준을 잡기 어렵다[] 표 1과 같은 문장들은 인간마다 느끼는 감정이 다를 수 있으므로 어떠한 한 감정으로 명확한 기준을 정하기 위해 설문조사를 진행하였다.

1.2 설문조사

감정을 명확하게 파악하기 어려운 100문장을 읽고 문장이 내포한다고 생각되는 감정을 선택하는 객관식형 설문조사를 진행하였다. 문장은 실험에서 모델을 학습하기 위한 라디오 청취자 문자 사연 4,000개 중 감정 분석이 어려운 100문장이며, 설문 응답을 통하여 인간의 보편적인 감정을 조사하였다.

표 2. 설문조사 정보
Table 2. Information of Survey

the number of respondents	163
type of answer	7 emotions 'neutrality', 'happiness', 'sadness', 'anger', 'surprise', 'hatred', 'fear'

설문 결과, 상황만 묘사된 문장의 감정은 '중립'을 선택해야 한다는 의견이 다수였다. 또한, 모호한 감정 표현은 응답자들이 가장 많이 선택한 감정을 기준으로 데이터셋을 라벨링하고, 실험을 진행하였다.

1.3 실험

본 논문에서는 데이터셋별 성능을 평가하기 위해 라디오 청취자 문자 사연으로 실험을 진행하였다. 두 경우의 정확도를 비교하기 위해 감정 클래스는 통일하였으며, 개방 데이터셋 '한국어 감정 정보가 포함된 단발성 대화 데이터셋'과 실제 라디오 청취자 문자 사연으로 각각 모델에 학습하여 문장을 테스트하였다. 개방 데이터셋은 온라인 댓글을 웹크롤링한 데이터셋으로 다양한 주제에 대한 본인의 의견 위주로 구성되어 있는 반면, 라디오 청취자 문자 사연은 본인의 감정 상태에 대한 문장이 많다는 차이점이 있다.

표 3. 실험 환경
Table 3. experiment environment

OS	Windows 11 64bit
Environment	Google Colaboratory
DeepLearning Library	Pytorch
Train Model	KoBERT

첫째로, 개방 데이터(38,594문장)로 모델을 학습하고, 라디오 청취자 문자 사연에 대한 감정 분석 테스트를 수행하였다. 둘째로, 라디오 청취자 문자 사연을 직접 수집하여 설문 결과를 바탕으로 감정 라벨링을 수행한 데이터셋(4,000문장)을 KoBERT 모델에 학습하고, 사용한 라디오 청취자 문자 사연 데이터셋을 제외한 라디오 청취자 문자 사연들에 대한 감정 분석을 수행하였다.

1.4 실험 결과

다음 표 4는 본 연구에서 두 종류의 데이터셋으로 KoBERT 모델을 학습하여 라디오 청취자 문자 사연에 대한 감정 분석을 수행한 뒤, 성능을 비교한 것이다.

표 4. 정확도
Table 4. Accuracy

Open Dataset		Radio Listeners' Messages	
Train accuracy	96.45%	Train accuracy	96.24%
Test accuracy	53.28%	Test accuracy	74.14%

데이터셋을 각각 학습하는 과정에서 Train accuracy는 약 96%로 차이가 거의 없었고, Test accuracy에서는 라디오 청취자 문자 사연을 활용한 경우가 개방 데이터셋을 활용했을 때보다 20.96% 높은 것을 볼 수 있었다. 두 데이터셋을 각각 학습하여 감정 분석을 수행한 결과, 개방 데이터셋을 학습한 감정 분석 모델은 60.42%의 정확도를 가졌고, 설문을 기반으로 라벨링한 라디오 청취자 문자 사연 데이터셋을 학습한 감정 분석 모델은 84.30%의 정확도를 가졌다. 데이터셋은 감정 라벨링에 대한 분류 기준의 명확성이 정확도에 큰 영향을 미치고, 테스트 문장과 비슷한 특성을 가진 데이터셋을 모델에 학습해야 정확도가 높아진다는 것을 알게 되었다. 본 연구에서도 라디오 청취자 문자 사연의 감정 분석을 수행할 때, 같은 특성을 가진 데이터셋(라디오 청취자 문자 사연)으로 학습한 경우의 정확도가 더 높았다.

개방 데이터는 시사에 대한 의견이 다수인 반면, 라디오 청취자 문자 사연은 본인의 감정 상태와 일상에 관한 이야기가 주를 이룬다. 이러한 특성 차이로 인하여 사용 단어가 달라 모델이 학습하는 방향이 달라졌다. 감정 분석 모델을 학습하는 데이터셋과 감정을 분석하고자 하는 데이터의 특성을 최대한 일치시켜 감정 분석 모델을 학습하여야 정확도가 높아진다. 또한, 상황과 문맥에 따라 달라지는 감정 표현이나 모호한 감정 표현에 대한 라벨링의 기준을 확실히 하여 인공지능이 혼돈을 느끼지 않고 감정을 정확히 분류할 수 있도록 데이터셋을 구성해야 한다.

Ⅲ. 결 론

본 논문에서 한국어 감정 분석을 기반으로 한 한국어 데이터셋에 대한 고찰과 한국어 감정 분석의 정확도 향상을 위한 설문조사와 실험을 진행하고 그 결과를 소개하였다. 한국어 감정에 대한 개방 데이터셋을 사용한 경우와 명확한 기준으로 감정을 분류하고 테스트 문장과 같은 특징을 가진 데이터셋을 사용한 경우를 비교했을 때, 후자의 경우가 감정 분석 결과 정확도가 더 높았다. 본 논문의 내용을 바탕으로 인공지능이 텍스트에서 인간이 느끼는 감정을 세부적으로 분류할 수 있도록 한국어 다중 감정 분석의 정확도를 향상하기 위한 자료로 쓰이는 데에 의미가 있다.

본 연구는 감정 클래스의 종류를 통일시켰으나 학습시킨 데이터셋의 개수가 다르다는 것과 모델이 복수 응답을 할 수 없어 한가지 감정으로만 분석했다는 한계가 있다.

추후엔 감정 분석 결과가 복합적인 감정이 드러나는 문장의 감정을 분석할 수 있는 시스템에 관해서 연구할 예정이다. 또한, 모호하고 복잡한 감정 표현과 비문법적 유행어를 사용한 감정 표현에 대한 데이터셋을 구성할 필요가 있고 이를 바탕으로 보다 정교한 감정 분석 딥러닝 모델을 개발하고자 한다.

참 고 문 헌 (References)

[1] Kim.Jihee, Oh.Jinhee, Kim.Myeungjin, Lim,Yankyu, "A Study on the Method of Creating Realistic Content in Audience-participating Performances using Artificial Intelligence Sentiment Analysis Technology", The Korean Society of Broadcast and Media Engineers, Vol.26, No.5, pp.533-541, 2021.
doi: http://doi.org/10.5909/JBE.2021.26.5.533

[2] Sudharsan Ravichandiran, Getting Started with Google BERT, (H. Jeon, S. Jung, H. Kim, Trans.), Hanbit Media, pp.22-74, 341-344, 2021

[3] Aurélien Géron, Hands-on Machin Learning with Scikit-Learn,Keras, and TensorFlow:Concepts,Tools,and Techniques to Build Intelligent Systems, (H. Park,Trans.), O'reily Media, , pp.598-670, 2020.

[4] Kwang-Hyeon Pak, Seung-Hoon Na, Jong-Hoon Shin, Young-Kil Kim, "BERT for Korean Natural Language Processing: Named Entity Tagging, Sentiment Analysis, Dependency Parsing and Semantic Role Labeling", Korea Computer Congress 2019, Korea, pp.584-586, 2019

[5] Yeonji Jang, Jiseon Choi, Hansaem Kim, "KcBert-based Movie review Corpus Emotion Analysis Using Emotion Vocabulary Dictionary" Journal of KIISE, Vol.49, No.8, pp.608-616, 2022.8.
doi: https://doi.org/10.5626/JOK.2022.49.8.608

[6] Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, Hyopil Shin,"A Small-Scale Korean-Specific BERT Language Model", Journal of KIISE, Vol. 47, No. 7, pp. 682-692, 2020. 7.
doi: https://doi.org/10.5626/JOK.2020.47.7.682