졸업논문
소개

라디오 청취자 문자 사연을 활용한 KoBERT 기반 한국어 다중 감정 분석 연구

이재아 / 서울과학기술대학교 Computer Vision Lab.

최근 딥러닝 기술 연구의 발전으로 감정 분석에 관한 다양한 연구가 진행되고 있다. 초기 자연어처리 분야에서는 인공지능이 인간의 감정 또는 감성을 단순 극성인긍/부정으로 분류하는 연구가 다수 존재하였다. 그러나 최근에는 긍/부정으로 감정 극성을 분류하는 이진 감성 분석을 넘어서 더 복잡하고 어려운 태스크인 다중 감정 분석에 관한 연구로 발전하고 있다.

이러한 다중 감정 분석 기술은 방송 분야와 융합하여 새로운 결과 창출을 기대할 수 있다. 그러나 방송 분야에서의 감정 분석 연구는 높은 관심에도 불구하고 아직 부족한 실정이다. 특히, 방송 매체 중 라디오에서 청취자 문자 사연은 실제 인간이 가질 수 있는 다양한 감정이 담겨 있는 방대한 양의 텍스트 데이터임에도 불구하고 관련 연구는 미흡할 뿐만 아니라 실제 사람들이 사용하는 문장에 대한 한국어 다중 감정 분석에 관한 연구는 더욱 필요하다. 이에 필자는 한국어 문장에서 다양한 감정을 추출하는 방법에 대해 고민하고, 실제 라디오 방송 환경에서 수집한 라디오 청취자 문자 사연을 활용하여 연구를 진행하였다. 본 연구를 진행하며 한국어 다중 감정 분석만의 고유한 특성과 감정 분석 모델의 성능을 높이고자 하는 방향성에 대해 고찰하였다.

본 연구에서는 실제 라디오 방송 환경에서 직접 수집한 한국어 데이터셋을 분석함으로써 한국어 다중 감정 분석이 어려운 이유에 대하여 고찰해 봤으며, KoBERT를 기반으로 구축한 한국어 다중 감정 분석 모델을 통하여 한국어 감정 분석의 정확도 향상을 위한 설문조사와 실험을 진행하고 그 결과를 소개하였다. 기존의 감정 분석 연구에서 보편적으로 이용한 개방 데이터셋이 아닌 실제 라디오 방송의 청취자 문자 사연을 직접 수집하여 감정 분석을 위한 한국어 데이터셋으로 활용했다는 점에서 차별성이 있으며, 실제 환경에서 수집한 라디오 청취자 문자 사연을 분석함으로써 한국어 감정 분석이 어려운 언어학적 특성에 대하여 고찰해 볼 수 있었다.

본 연구에서는 한국어 다중 감정 분석의 정확도를 높일 수 있는 데이터셋 구성에 관한 고찰과 분석을 위해 설문조사와 두 가지 실험을 수행하였다. 실험을 진행하기에 앞서, 모델을 학습하기 위한 한국어 말뭉치를 구축하고 감정 레이블링의 보편적인 기준을 정의하기 위하여 163명을 대상으로 한국어 문장에 대한 설문조사를 진행하였다. 여기서 한국어 말뭉치란, 실제 인간이 사용한 한국어 문장에 감정 정

보를 부착한 텍스트 데이터셋을 의미한다. 설문조사를 통하여 인간이 보편적으로 느끼는 감정에 대해 정의하고, 모델이 인간의 감정을 학습하는 방향성을 제시할 수 있었다. 이를 통하여 실험에 사용될 한국어 말뭉치를 직접 구축하여 실험에 활용하였다.

한국어 및 문어체에 특화된 KoBERT 언어 모델로 한국어 다중 감정 분석 시스템을 구축하여 두 가지 실험을 진행하였다. 실험에서 사용한 KoBERT 모델은 BERT라는 모델을 한국어에 맞게 개선한 버전으로서 SKT Brain에서 공개한 사전학습 기계 번역 언어 모델이며, 한국어 감정 분석 관련 연구에서 뛰어난 성능을 입증하였다. 두 가지 실험에서 KoBERT 모델이 라디오 청취자 문자 사연에 대하여 '행복', '슬픔', '놀람', '분노', '공포', '혐오', '중립' 총 일곱 가지 감정으로 분석을 수행한다. 첫 번째 실험에서는 같은 한국어 말뭉치로 전이학습(Fine-tuning)을 수행하고, 실제 환경에서 수집한 라디오 청취자 문자 사연을 정제한 데이터와 정제하지 않은 데이터에 대하여 감정 분석을 수행하여 비교 분석하였다. 이를 통하여 KoBERT 모델은 유행어, 오타 등 비문법적인 요소의 존재 여부에 상관없이 한국어 다중 감정 분석에 뛰어난 성능을 보이는 것을 알게 되었다. 두 번째 실험에서는 AI HUB에서 제공하는 개방 데이터셋 '한국어 감정 정보가 포함된 단발성 대화 데이터셋'과 설문조사 결과를 참고하여 라디오 청취자 사연에 감정 정보를 직접 부착한 한국어 말뭉치를 각각 전이학습을 위한 데이터셋으로 사용하여 라디오 청취자 문자 사연에 대한 감정 분석을 수행하였다. 이를 통하여 한국어 다중 감정 분석 모델의 성능 향상을 위한 전이학습용 데이터셋 구성 방안을 제안하였다. 실험 결과, 같은 환경에서 수집되어 같은 특성을 가진 데이터셋을 사용한 경우가 전이학습 과정에서 테스트 정확도부터 차이가 나기 시작하여 감정 분석 결과 정확도도 더 높은 것을 알 수 있었다. 이는 감정 레이블링에 대한 기준의 명확성이 감정 분석 시스템의 성능에 큰 영향을 미치고, 모델에 주입하는 테스트 문장과 비슷한 특성을 가진 데이터셋을 모델에 학습해야 정확도가 높아진다는 것을 의미한다.

본 연구의 내용을 바탕으로 인공지능이 텍스트에서 인간이 느끼는 감정을 세부적으로 분류할 수 있도록 한국어 다중 감정 분석의 정확도를 향상하기 위한 자료로 쓰이는 데에 의미가 있다. 본 연구를 구체화한다면 실제 방송 환경에도 활용이 가능할 것으로 보이며, 방송 통신 융합 분야의 발전을 위한 기반이 될 수 있을 것이다.

이 재 아

- 2016년 : 서울과학기술대학교 전자IT미디어공학 전공 학사
- 2023년 : 서울과학기술대학교 나노IT디자인융합대학원 정보통신미디어공학 전공 석사
- 2017년 ~ 현재 : 가톨릭평화방송 재직 중
- 주관심분야 : 라디오방송, 차세대방송기술, 감정분석, 딥러닝