멀티모달 데이터에 대한 특성공학 기반의 복합 모델을 활용한 발화 감정 분석 기술

Emotion Recognition in Conversation Using Ensemble Model based on Feature Engineering About Multi-Modal Data

요 약

본 연구에서는 멀티모달 감정 데이터셋인 KEMDy20 데이터셋에 포함된 스크립트 데이터, 음성 데이터, 바이오 데이터를 복합적으로 활용하여 감정 분석의 정확도를 높이는 것을 목표로 한다(1). 자연어 처리 모델을 활용하여 스크립트 데이터로부터 발화의 긍·부정 점수를 산출하고(2-1), 이를 참고하여 음성 데이터의 MFCC 분석을 통해 감정 분석을 진행하였다(2-2). 바이오 데이터에 존재하는 문제점을 특성공학을 이용하여 처리한 후, CNN과 LSTM을 결합한 새로운 모델을 통해서 감정 분석을 진행하였다(2-3). 본 연구에서 제안한 특성공학과 모델을 이용하면 감정 분석에서 더 나은 결과를 도출할 수 있다(3).

1. 서 론

Emotion Recognition in Conversation(이하ERC)은 사람의 대화에서 음성, 생체 신호, 각종 데이터를 추출하여 인간이 느끼는 감정을 분석하는 기술이다. 최근 ERC는 대화형AI, 소비자 만족도 조사, 개인 맞춤형 환경제공 및 서비스 등 여러 분야에서 활용되고 있는 추세이다. 많이 활용되는 만큼 감정 분석의 정확성을 높이는 것은 해당 연구 분야의 주된 목표이다. 또한, 현재까지의 연구를 보면 딥러닝을 이용한 음성 감정 인식 기술이나 EEG(Electroencephalogram)신호를 Recurrent Neural Network(RNN)모델로 학습시킨 감정인식 기술 등의 연구가 진행되었다[1-2].

하지만 음성이나 EEG신호 같은 개별의 데이터만으로 감정을 인식하는 것은 멀티모달 데이터로 감정을 인식하는 것보다 정확도가 높지 않다.

따라서 본 논문에서는 한국어 기반 멀티모달 감정 데이터셋(KEMDy20; Korean Emotional Multi-modal Dataset in 2020)의 텍스트, 음성, 바이오 데이터를 Word2vec와 MFCC(Mel-Frequency Cepstral Coefficient), ERCL(Emotion Recognition CNN-LSTM)을 합친 특성공학으로 감정 분류의 정확도를 개선하는 방법을 제시하고자 한다.

본론1에서는 대화 스크립트를 이용하여 감정을 분류한다. 텍스트의 특성상 같은 단어지만 문맥상 의미가 달라지는 단어들이 있다. 이를 보완하기 위해서 감정에 대한 점수가 포함된 파일을 가져오고, 유사한 의미의 문장을 찾아주는 Word2vec로 텍스트 데이터와 각각 매칭시켜준다. 매칭 시킨 텍스트마다 감정 점수가 매겨지고 이 점수를 계산하여 각 문장에 대한 긍정도와 부정도를 파악한다. 추가로 감정의 분류를 좀 더 정확하게 하기 위해서 음성 데이터와 바이오 데이터를 이용한 AI시스템을 구축하였다.

본론2에서는 음성 데이터를 이용한 감정 분석을 수행한다. 음성 데이터 분석에 유용한 MFCC는 음성으로부터 특징벡터를 추출하여 2차원 형태의 배열로 나타낸다. MFCC를 이미지처럼 생각하여 Convolutional Neural Network(CNN) 모델로 학습시켜 감정을 7가지 레이블(neutral/ happy/ angry/ sad/ fear/ surprise/ disgust)로 분류한다.

본론3에서는 바이오 데이터를 이용해서 감정 분석에서의 Accuracy를 높일 수 있는 방법을 모색하였다. KEMDy20의 바이오 데이터는 E4센서를 통해서 측정한 사람의 생체 신호를 시간에 따라 데이터로 변환한 시계열 데이터이다. 하지만 KEMDy20의 바이오 데이터는 결측치가 존재하기 때문에 결과를 예측하기 위해 필요한 데이터가 충분하지 않다. 그러므로 해당 문제를 해결하기 위하여 시계열 데이터를 다룰 때 유용한 LSTM모델을 통해 결측치를 채워준다. 또한, 감정 분류의 정확도를 높이기 위해서 LSTM모델에 CNN 모델을 추가한 ERCL모델을 설계한다.

2. 본 론

2-1. Word2vec을 이용한 스크립트 분석

데이터 파일의 wav 디렉토리 안의 각 세션에는 발화자의 스크립트 데이터가 텍스트 파일로 정리되어 있다. 이 텍스트로 된 대사들을 통해 긍정과 부정의 의미를 예측하는 것이 이 장의 목표이다. 감성 점수를 부여하기 위해서 naver sentiment movie corpus v1.0(이하 nsmc)를 사용한다. 여기서 사용하는 nsmc는 네이버에서 제공하는 15만개의 네이버 영화리뷰 데이터이다. 리뷰들과 리뷰에 대한 긍정, 부정 점수를 모아 놓았기 때문에 이 데이터와 텍스트 파일을 연관 분석하여 긍정, 부정 점수를 추출하였다. 이를 위해서 문장을 벡터화하여 유사도 분석을 수행하는 Word2vec을 사용한다.

Word2vec가 학습하기 위해서 nsmc 데이터에 존재하는 NaN값들을 모두 제거하고 한국어 자연어 처리기인 KoNLPy의 Okt 형태소 분석으로 토큰화 후 모델을 학습시켰다. 모든 텍스트 파일에 똑같은 방식으로 토큰화를 진행한 후 각 토큰에 대해 모델에서 임베딩 벡터를 가져온다. 모든 벡터를 추출하면 이를 통해 평균값으로 문장의 벡터를 계산한다. 그리고 문장 벡터와 코사인 유사도를 계산하여 가장 벡터값이 유사한 문장을 nsmc에서 가져온다. 코사인 유사도를 사용하면 의미가 가장 유사한 문장을 가져올 수 있다. 마지막으로 그 문장이 nsmc에서 저장된 감정 점수를 불러올 수 있도록 하였다.

위 과정을 수행하면 유사한 문장이 긍정의 의미를 나타낼 때 1의 값을 적용하고 부정의 의미를 나타낼 때 0의 값을 적용한다. 비슷한 의미의 문장을 찾지 못해 의미를 갖지 않는다면 unknown 값을 적용한다. 충분한 양의 데이터로 만든 모델을 사용해 텍스트 파일에서 단어들의 임베딩 벡터를 계산하고 문장 벡터를 만들어 가장 유사한 문장을 추출하는 과정을 진행하기 때문에 스크립트에서 부정과 긍정의 감정 예측 결과를 가져올 수 있다. 이 결과로 이후의 MFCC와 연관성을 분석하여 감정 예측에서의 효과를 볼 수 있다.

2-2. 음성 데이터를 통한 감정 분석

음성 데이터 분석 및 처리에 있어서 가장 대표적인 방법은 MFCC를 이용한 방법이다[3-5]. MFCC는 음성 데이터를 특징벡터화 해주는 알고리즘으로 MFCC에서 어떠한 특징을 뽑아내는가에 따라 음성 데이터 분석 성능에 크게 영향을 미친다. 따라서 적절한 특징벡터를 추출하는 것이 MFCC 성능에 중요한 문제라고 할 수 있다. MFCC는 일반적으로 13개의 특징벡터를 추출하지만 더 많은 특징벡터를 추출함으로써 다양한 정보를 얻을 수 있다. MFCC로 특징벡터화된 음성 데이터를 7가지의 감정 레이블로 분류하기 위해 CNN을 사용하였다[6]. CNN은 일정한 크기의 입력값을 받으므로 서로 다른 크기를 지닌 MFCC를 동일한 크기로 맞출 필요가 있다. MFCC의 크기를 맞추는 방법에는 Zero-Padding 방법과 Time-Stretching 방법이 있다. 본 논문에서는 MFCC에서 추출할 특징 벡터의 개수와 MFCC의 크기를 일정하게 맞추는 방법을 조합·비교하여 음성 감정 분석의 결과를 가장 잘 나타낼 수 있는 특성을 찾고자 한다. 아래의 그림1은 프로세스의 전 과정을 도식화한 그림이다.

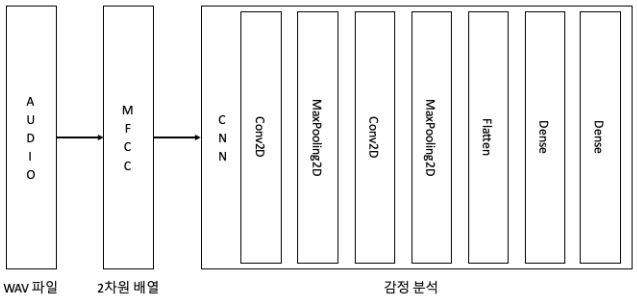

그림1. CNN 모델 구성 (MFCC-CNN)

wav 파일로 주어진 샘플 데이터 13462개에서 학습을 위한 학습 데이터 10769개와 Accuracy 측정을 위한 2693개 데이터로 나누었다. 테스트 데이터에는 감정 레이블이 골고루 분포할 수 있도록 하였다. 학습 데이터는 CNN을 사용하여 10번 이상 학습하였다. 학습과 테스트에 대한 최종 결과는 표1과 같다.

Data Preprocessing	Test Loss	Test Accuracy
Zero-Padding + N=13	0.9960	0.8336
Zero-Padding + N=22	1.3676	0.8232
Zero-Padding + N=31	2.6865	0.8236
Zero-Padding + N=40	1.9793	0.8473
Time-Stretching + N=13	0.8860	0.8251
Time-Stretching + N=22	1.0273	0.8265
Time-Stretching + N=31	1.3583	0.8069
Time-Stretching + N=40	1.5009	0.8057

표1. Loss & Accuracy

가장 우수한 성능을 보인 특성공학은 Zero-Padding + N=40으로 Loss의 값은 가장 높지만, Accuracy 부분에서 84.7%의 정확도를 보였다. 따라서 Zero-Padding + N=40을 통한 특징 추출 및 가공을 음성 감정 분석에서의 특성공학 방법으로 제시하고자 한다.

2-3. 바이오 데이터를 활용하기 위한 ERCL 설계

시계열 데이터에는 LSTM 모델을 사용하여 예측을 진행한다[7-8]. 서론에서 서술했던 LSTM의 단점을 보완하기 위해 CNN와 결합된 ERCL모델을 사용한다[9].

KEMDy20에 'IBI(Inter-beat Interval; 맥박 간격)' 컬럼에는 결측 데이터가 많이 있고, 이는 서론에서 언급했듯이 학습 데이터로 사용하기엔 큰 문제가 있다. 이런 문제를 해결하기 위해 LSTM 모델을 사용하여 'IBI'값 예측을 진행한다. 해당 과정을 통해 그림 2와 같이 'IBI'를 채워 넣는 과정을 거쳐 데이터셋을 완성했다. 이때 레이블로 사용될 'IBI' 값이 너무 적어 학습이 되지 않은 데이터는 삭제하여 총 51개의 데이터 파일을 만들었다. 각각의 파일을 8 : 2의 비율로 학습 데이터와 테스트 데이터로 나누어 Accuracy를 측정한다. 이를 통해 약간의 Accuracy가 개선되었다.

그림 2. IBI 결측치 예측

본 논문에서는 LSTM과 ERCL의 정확도를 비교하고자 한다. LSTM의 모델과 ERCL모델은 그림3에서 나타난다. 이를 통한 예측 결과는 표2와 같다. 표2는 두 가지 모델의 Accuracy에 관한 표이다.

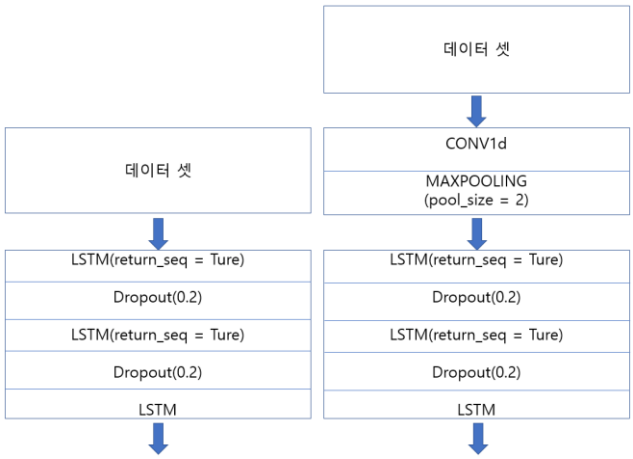

그림 3. LSTM과 ERCL

Model	Accuracy
LSTM	0.8233
ERCL	0.8267

표2. LSTM과 ERCL의 Accuracy 비교

3. 결 론

Word2vec, MFCC, ERCL의 특성공학의 결과를 활용하여 감정 분석을 진행하였을 때, 더 나은 Accuracy를 보였다.

Word2vec을 통해 나온 스크립트의 긍·부정 점수를 참고하여 음성 파일에 대한 분석을 실시하였다. MFCC에서 40개의 특징벡터를 추출하고 Time-Stretching 방법을 사용하여 크기를 맞췄을 때가 기존 방법의 평균 정확도인 82.9%보다 1.8%p 높은 84.7%의 정확도를 보였다. 바이오 데이터에서 IBI 결측치에 대한 특성공학 기법을 이용한 결과를 활용하여 감정 분석을 진행하였을 때, 기존의 데이터를 활용한 결과보다 0.34%p 높은 82.67%의 정확도를 보였다. 두 결과를 복합적으로 활용하면 더 나은 결과를 보일 수 있음을 확인하였다.

참고 문헌

[1] M.ChowdaryKalpana, J.Anitha, & D.HemanthJude. (2022). Emotion Recognition from EEG Signals Using Recurrent Neural Networks. "MDPI Open Access Journals".

[2] RUHULKHALILAMIN, EDWARDJONES, MOHAMADBABARINAYATULLAH, TARIQULLAH JAN, MOHAMMADZAFARHASEEB, & THAMERALHUSSAIN4. (2019). Speech Emotion Recognition Using Deep Learning Techniques: A Review. "IEEE Access", 117327~117345.

[3] 최하나, 변성우, & 이석필. (2015). 음성신호기반의 감정분석을 위한 특징벡터 선택. "The Transactions of the Korean Institute of Electrical Engineers Vol. 64", (페이지: 1363 ~ 1368).

[4] 윤상혁, 전다윤, & 박능수. (2021). CNN - LSTM 모델 기반 음성 감정인식. "ACK 2021 학술발표대회 논문집", (페이지: 939-941).

[5] 양종열, & 김흥국. (2007). MFCC 특징벡터를 이용한 감정인식 방법의 성능 비교., (페이지: 633-634).

[6] 신경식, 유신우, & 오혁준. (2020). MFCC와 CNN을 이용한 저고도 초소형 무인기 탐지 및 분류에 대한 연구. "한국정보통신학회논문지 Vol. 24", (페이지: 364~370).

[7] 이여진, 황동현, 이슬아, 조주필, & 고경석. (22). 다변수LSTM딥러닝네트워크를이용한육계시세예측모델연구. "The Journal of Korean Institute of Communications and Information Science", (페이지: 2058~2064).

[8] 박세희, 정의손, 박승보, & 박지영. (2021). LSTM 기반의 Handheld 샷 검출. "한국컴퓨터정보학회 하계학술대회 논문집 제29권 제2호", (페이지: 193~194).

[9] 김민기. (2022). 시분할CNN-LSTM기반의시계열진동데이터를이용한회전체기계설비의이상진단. "Journal of Korea Multimedia Society ", (페이지: 1547-155).