The Transactions of the Korea Information Processing Society
TKIPS Vol.13, No.4, pp.174~180

eISSN: 3022-7011
https://doi.org/10.3745/TKIPS.2024.13.4.174

Efficient Emotion Classification Method Based on Multimodal Approach Using Limited Speech and Text Data

Mirr Shin† · Youhyun Shin††

ABSTRACT

In this paper, we explore an emotion classification method through multimodal learning utilizing wav2vec 2.0 and KcELECTRA models. It is known that multimodal learning, which leverages both speech and text data, can significantly enhance emotion classification performance compared to methods that solely rely on speech data. Our study conducts a comparative analysis of BERT and its derivative models, known for their superior performance in the field of natural language processing, to select the optimal model for effective feature extraction from text data for use as the text processing model. The results confirm that the KcELECTRA model exhibits outstanding performance in emotion classification tasks. Furthermore, experiments using datasets made available by AI-Hub demonstrate that the inclusion of text data enables achieving superior performance with less data than when using speech data alone. The experiments show that the use of the KcELECTRA model achieved the highest accuracy of 96.57%. This indicates that multimodal learning can offer meaningful performance improvements in complex natural language processing tasks such as emotion classification.

Keywords : Artificial Intelligence, Natural Language Processing, Speech Recognition, Multimodal, Emotion Classification

적은 양의 음성 및 텍스트 데이터를 활용한 멀티 모달 기반의 효율적인 감정 분류 기법

신 미 르† · 신 유 현††

요 약

본 논문에서는 wav2vec 2.0과 KcELECTRA 모델을 활용하여 멀티모달 학습을 통한 감정 분류 방법을 탐색한다. 음성 데이터와 텍스트 데이터를 함께 활용하는 멀티모달 학습이 음성만을 활용하는 방법에 비해 감정 분류 성능을 유의미하게 향상시킬 수 있음이 알려져 있다. 본 연구는 자연어 처리 분야에서 우수한 성능을 보인 BERT 및 BERT 파생 모델들을 비교 분석하여 텍스트 데이터의 효과적인 특징 추출을 위한 최적의 모델을 선정하여 텍스트 처리 모델로 활용한다. 그 결과 KcELECTRA 모델이 감정 분류 작업에서 뛰어난 성능이 보임을 확인하였다. 또한, AI-Hub에 공개되어 있는 데이터 세트를 활용한 실험을 통해 텍스트 데이터를 함께 활용하면 음성 데이터만 사용할 때보다 더 적은 양의 데이터로도 더 우수한 성능을 달성할 수 있음을 발견하였다. 실험을 통해 KcELECTRA 모델을 활용한 경우가 정확도 96.57%로 가장 우수한 성능을 보였다. 이는 멀티모달 학습이 감정 분류와 같은 복잡한 자연어 처리 작업에서 의미 있는 성능 개선을 제공할 수 있음을 보여준다.

키워드 : 인공지능, 자연어 처리, 음성 인식, 멀티모달, 감정 분류

1. 서 론

본 논문은 wav2vec 2.0 모델을 활용하여 음성 인식을 통한 감정 분류의 성능을 향상시키는 방안을 연구한다. [1]에서는 제한된 데이터 샘플링이 음성 인식 성능에 미치는 영향에 초점을 맞추었으며, wav2vec 2.0 모델의 효율적 활용 방안을 모색하였다. 특히, 데이터 샘플링 기법이 모델의 성능에 어떠한 영향을 미치는지 분석하는 데 중점을 두었다.

이러한 기존 연구[1]를 바탕으로 더 발전시켜, 하나의 특징을 가지는 데이터만 학습하지 않고 서로 다른 특징을 가지는 데이터를 학습하는 멀티모달 학습을 적용하여 성능의 변화를 관찰하였다. 멀티모달 학습은 다양한 유형의 데이터 예를 들어, 음성과 텍스트, 텍스트와 영상 혹은 세 가지 전부를 사용하는 등 서로 다른 특징을 가진 데이터를 통합하여 사용함으로써, 더 풍부하고 다차원적인 정보를 활용할 수 있다. 이는

※ 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학석사연계ICT 핵심인재양성사업의 연구결과로 수행되었음(IITP-2024-RS-2023-00260175).
※ 이 논문은 2023년 ACK 2023의 우수논문으로 "wav2vec2.0을 활용한 한국어 음성 감정 분류를 위한 데이터 샘플링 전략"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 인천대학교 컴퓨터공학과 석·박사통합과정
†† 정 회 원 : 인천대학교 컴퓨터공학부 부교수
Manuscript Received : December 26, 2023
First Revision : March 5, 2024
Accepted : March 18, 2024
* Corresponding Author : Youhyun Shin(yhshin@inu.ac.kr)

Table 1. Statistics of Participants and Data Quantity by Emotion

Label	Num. of Participants	Num. of Data	Num. of Total Data
Happy	41	66,097	453,365
Anxious	40	65,853	
Angry	40	65,715	
Sad	40	64,359	
Embarrassed	40	64,201	
Neutrality	40	63,400	
Hurt	40	63,740	

음성 인식을 통한 감정 분류의 정확도 향상에 기여할 수 있다. 음성 데이터뿐만 아니라 텍스트 데이터도 같이 활용하여 실험을 진행하였고 감정 분류의 성능이 향상되는지 확인하였다.

음성 데이터와 함께 텍스트 데이터를 활용하는 멀티모달 학습 방법이 감정 분류의 성능을 유의미하게 향상시킬 수 있다고 생각하였다. 이는 후술한 텍스트 데이터만을 활용하여 감정 분류를 진행한 실험에서 확인할 수 있다. 사용할 데이터 세트[1]는 같은 대본을 가지고 여러 명의 화자가 읽는 형식으로 음성 데이터에 비해 텍스트 데이터는 중복이 존재하여 양이 상대적으로 적은 데이터이다. Table 1은 데이터 세트에 감정마다 녹음 화자 수와 데이터양을 보여준다.

전체 데이터는 453,365개의 데이터이고 이는 음성과 텍스트 데이터를 모두 포함하고 있다. 데이터 세트에서 음성과 텍스트를 모두 제공하고 이를 전부 활용하는 방법이 감정 분류의 성능을 높일 수 있다고 생각하였다. 이를 확인하기 위해서 2가지 실험을 먼저 진행하였다.

첫 번째, 텍스트 데이터만 사용하여 감정 분류를 진행하여 텍스트 데이터만을 사용하여 성능을 비교할 기준을 확인하는 실험이 필요하다.

두 번째, 텍스트 데이터의 성능과 기준을 확인하였다면 텍스트 데이터의 특징을 더욱 잘 끌어내 줄 모델을 찾을 필요가 있다.

이후 음성과 텍스트 데이터의 특징을 모두 결합하여 학습하는 멀티모달 학습으로 감정 분류를 진행한다. 이러한 실험은 멀티모달 학습이 음성 인식 시스템의 데이터를 효율적으로 처리하여 더욱 정확도 높은 감정 분류 모델을 제안한다.

2. 관련 연구

2.1 음성 처리 모델

본 논문에서는 음성 데이터의 특징을 추출하기 위해 wav2vec 2.0[2] 모델을 활용하였다. wav2vec 2.0은 페이스북 AI에 의해 개발된 모델로 음성 인식 분야에 많이 활용되는 대표적인 모델이다. 이 모델은 원시 음성 데이터에서의 비지도 학습에 초점을 맞추어 wav2vec[3]을 발전시킨 모델이다. 특히 wav2vec 2.0은 고급 자연어 처리 기술과 결합하여, 음성 데이터에서 직접 학습할 수 있는 강력한 기능을 제공한다. 이 모델은 대규모 음성 코퍼스에서 효과적으로 특징을 추출하며, 이를 통해 음성 인식, 음성 기반 감정 분석 등 다양한 분야에서 뛰어난 성능을 보여준다.

wav2vec 2.0의 핵심은 컨볼루션 신경망을 사용하여 음성의 원시 파형을 인코딩하고, 이후 Transformer[4] 기반의 아키텍처를 통해 더 높은 수준의 음성 특징을 학습하였다. 이 과정에서, 모델은 데이터의 표현을 더욱 잘 학습하였다.

이러한 특성 덕분에 wav2vec 2.0은 음성 인식 분야에서의 새로운 가능성을 열어주며, 특히 다양한 언어를 처리하는 데 있어서 높은 유연성과 정확성을 제공한다. 이 논문에서는 이러한 점을 토대로 좋은 음성 특징과 텍스트를 활용하여 적은 데이터로도 좋은 성능을 이끌어내었다.

2.2 텍스트 처리 모델

텍스트 데이터의 특징을 효과적으로 추출하기 위해 어떤 한국어 모델이 좋을지 실험을 통해 성능을 비교하였다. 본 논문에서는 BERT[5], RoBERTa[6], ELECTRA[7] 3개의 모델을 비교할 예정이고, 추가로 한국어로 학습한 모델의 성능을 비교하였다.

BERT는 자연어 처리 분야에서 혁신적인 변화를 가져온 모델 중 하나이다. BERT는 Transformer의 인코더 부분을 사용하여 문장 내의 모든 단어를 한 번에 고려함으로써, 문맥에 따른 단어의 의미를 더 정확하게 파악할 수 있게 한다. 이는 이전의 단방향 모델들이 가진 한계를 극복한 것으로, BERT는 텍스트 분류, 의미론적 유사성 평가, 질의응답 시스템 등 다양한 자연어 처리 작업에서 뛰어난 성능을 보였다.

RoBERTa는 BERT의 한계를 극복하기 위해 개발된 모델로, 더 많은 데이터와 동적 마스크, 더 긴 학습 시간, 더 큰 배치 크기를 사용하여 BERT의 사전학습 과정을 최적화한다. RoBERTa는 동일한 아키텍처를 유지하면서 학습 과정에서의 여러 변화를 통해 BERT보다 더 나은 성능을 달성했다. 이러한 최적화는 모델이 더 넓은 범위의 문맥과 더 복잡한 언어 패턴을 이해할 수 있게 하여, 다양한 자연어 처리 분야에서 성능을 개선하였다.

ELECTRA는 BERT와 RoBERTa의 발전된 아이디어를 바탕으로 하면서도, 대체 토큰 탐지라는 새로운 학습 방식을 도입함으로써 또 다른 차별점을 제시한다. 기존의 모델이 문맥을 이해하기 위해 다음 단어를 예측하는 방식을 사용했다면, ELECTRA는 문장 내 일부 단어를 인공적으로 생성된 토큰으로 대체하고, 이를 원래의 토큰과 구별하도록 학습한다. 이 방식은 효율성과 성능 면에서 특히 작은 크기의 모델에서도 뛰어난 결과를 보여준다. ELECTRA의 이러한 접근 방식은 전체 문장의 맥락을 더 잘 이해하고, 더 정교한 언어 이해 능력을 보였다.

1) https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=466

2.3 멀티모달

최근 자연어 처리 및 음성 인식 분야에서 멀티모달 학습 방법이 주목받고 있다. [8-11]과 같은 연구들을 통해 한 유형의 데이터만 사용하기보다 영상, 오디오, 텍스트 등 여러 유형의 데이터를 사용하는 것이 효과적임을 알 수 있다. 이러한 학습 방법은 각각의 음성과 텍스트 데이터가 가진 정보의 다양성과 풍부함을 활용하여, 더 정확하고 효과적인 결과를 도출할 수 있다. 한국어 데이터를 활용한 연구도 활발하게 진행되고 있다. 한국어 데이터 관련 연구를 좀 더 자세히 보자면 [12]는 SpeechT5를 활용하여 wav2vec 2.0과 SpeechT5를 각자 사용할 때와 함께 사용하는 멀티모달 학습의 성능을 평가하였다. [13]은 추출된 특징 벡터를 앙상블하여 성능을 평가하였다. [14]는 Wav2vec 2.0과 한국어로 학습된 RoBERTa 모델을 사용하여 성능을 평가하였다. [15]는 Contextualized GNN을 활용하여 성능을 평가하였다. 지난 연구들은 추출된 특징을 결합하는 방식에 중점을 두고 실험을 진행하였다. 우리가 진행한 실험에서는 이러한 특징을 추출하는 최적의 모델을 찾는 실험을 진행하였다. 또한, 효율적인 데이터 활용을 위한 샘플링 기법이 추가되었다.

텍스트와 음성 데이터를 활용한 멀티모달 학습은 음성 인식을 통한 감정 분류의 정확도를 향상시키고, 더 복잡한 언어 이해 작업을 수행할 수 있음을 보여준다. 예를 들어, 음성 데이터의 톤이나 강조와 같은 요소와 텍스트 데이터의 문맥적 정보를 결합함으로써, 더 정교한 의미 분석이 가능해진다.

3. 본 론

비교를 위해 [1]의 실험 방법을 참고하여 데이터를 양과 길이로 샘플링하여 실험을 진행한다. (4.1)

음성 데이터의 성능과 샘플링 방법을 확인하고 이후 텍스트와 멀티모달 학습 실험을 진행한다. (4.2)

실험의 첫 단계에서는 텍스트 데이터만을 사용하여 감정 분류 모델의 성능을 평가한다. 이를 통해 모델이 텍스트 데이터만을 활용했을 때의 기준을 설정하는 것이 목표이다. (4.3)

이후, 실험의 두 번째 단계에서는 멀티모달 학습을 적용하여 성능 평가를 수행한다. 멀티모달 학습은 감정 분류 시스템이 두 가지 유형의 데이터를 동시에 처리하고, 각 데이터에서 특징을 추출하고 두 특징을 합쳐 성능을 향상시키는 것을 목표로 한다. 실험 결과는 음성 데이터 또는 텍스트 데이터만을 사용했을 때와 비교하여 멀티모달 학습이 얼마나 감정 분류의 성능을 향상시킬 수 있는지 중점을 두고 분석한다. (4.5)

추가로 텍스트 데이터의 특성을 고려하여 실험을 진행한다. 음성 데이터에 비해 텍스트 데이터에는 중복되는 부분이 많아, 상대적으로 적은 양의 데이터로 실험을 진행한다. 실험을 통해서 텍스트 데이터는 음성 데이터보다 양이 적어도 더 좋은 성능을 보여주고 더 많은 표현을 학습할 수 있다. 또한, 그러한 학습을 위한 표현을 최대한으로 끌어내기 위해 최적의 텍스트 처리 모델들의 비교 실험을 진행한다. (4.4)

이러한 실험을 통해, 텍스트 데이터가 감정 분류의 성능에 미치는 영향과 멀티모달 학습이 성능 향상에 기여하는 정도를 파악할 수 있다. 이는 음성 인식 감정 분류 시스템의 효율적인 데이터 활용 방안과 멀티모달 학습의 중요성을 강조하는 결과다. 텍스트 데이터의 활용을 통해 성능이 향상되는 것을 보아 감정 분류에서는 음성과 텍스트를 함께 사용하는 학습이 좋다는 결과를 확인할 수 있다.

Fig. 1은 전체적인 실험 진행도이다. 이 실험은 멀티모달 학습을 통해서 음성과 텍스트 데이터의 모든 특징 정보를 활용하는 과정을 담고 있다. Step 1에서는 음성 처리 모델이 음

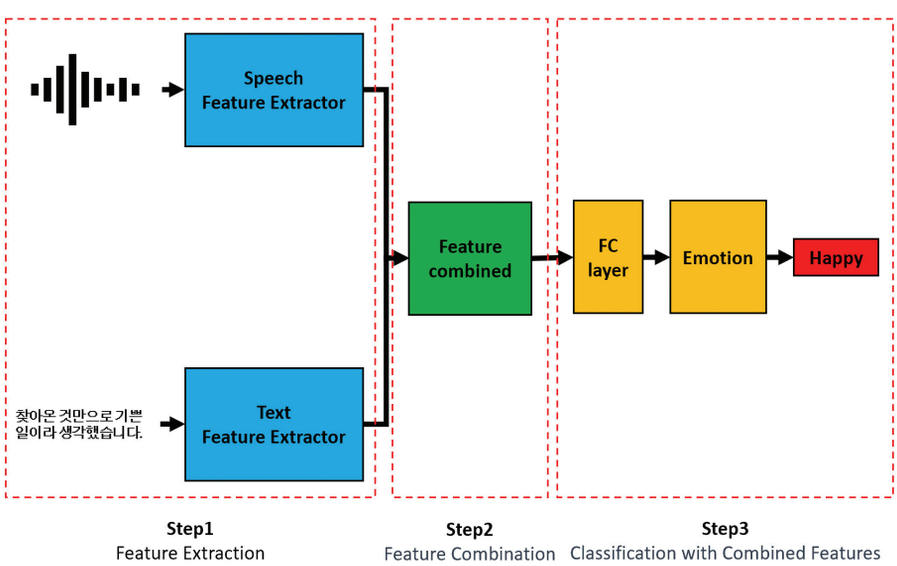


Fig. 1. Overview of the Experimental Process Depicting Step 1: Feature Extraction, Step 2: Feature Combination, and Step 3: Classification Using Combined Features

성 데이터로부터 특징 벡터를 추출하고, 텍스트 처리 모델은 텍스트 데이터에서 특징 벡터를 추출한다. 이 두 모델은 각각의 데이터 유형에서 중요한 정보를 포착하여 특징 벡터로 변환하는 역할을 한다. Step 2에서는 Step 1에서 추출된 음성과 텍스트의 특징 벡터들을 결합하여 사용한다. 이 과정을 통해 결합한 특징 벡터는 음성과 텍스트 데이터의 정보를 통합적으로 반영하는 새로운 특징 벡터가 된다. Step 3 감정 분류 마지막 단계에서는 결합한 특징 벡터가 FC(Fully Connected Layer)를 통과한다. 이 FC는 결합한 특징 벡터를 바탕으로 최종적인 감정 분류를 수행한다. 이 단계에서는 결합한 특징 벡터의 통합적인 분석을 통해 감정을 분류하는 것이 목표이다.

아래는 음성과 텍스트 데이터에서 특징을 추출한 이후 과정을 수식화한 것이다. 음성 데이터의 특징을 A, 텍스트 데이터의 특징을 T라 하고 이 두 특징을 결합하여, 멀티모달 특징 벡터 x를 생성한다[Equation (1)].

$$x = concatenate(A, T) \tag{1}$$

생성된 x 벡터를 첫 번째 선형 계층을 통과시키고 ReLU 활성 함수를 적용하여 h 벡터를 얻는다[Equation (2)].

$$h = ReLU(W_1 x + b_1) \tag{2}$$

중간 표현 h를 두 번째 선형 계층을 통화시켜, 감정 분류를 위한 최종 벡터 y를 생성한다[Equation (3)].

$$y = W_2 h + b_2 \tag{3}$$

벡터 y에 softmax 함수를 적용하여, 각 감정 클래스에 속할 확률을 분포를 얻는다. 최종적으로 각 입력에 대한 감정 분류 결과를 얻는다[Equation (4)].

$$\hat{e} = softmax(y) \tag{4}$$

4. 실 험

실험에 사용한 GPU는 NVIDIA A100 PCIe 40GB GPU를 사용하였다. batch size는 32, epoch은 10으로 통일하여 실험을 진행하였다. 사용한 음성 처리 모델은 huggingface의 facebook/wav2vec2-large-robust로 음성 데이터를 텍스트 데이터는 4.4절에서 비교한 모델들 중 제일 성능이 좋았던 beomi/KcELECTRA-base-v2022 모델을 사용하여 실험을 진행하였다.

4.1 데이터 샘플링

비교를 위해 두 가지 방법으로 샘플링을 진행하였다. 데이터의 수와 데이터의 길이를 기준으로 데이터를 샘플링하여 실험을 진행하고 이에 따른 성능의 변화를 관찰하였다. 이러한 샘플링을 통해 데이터의 길이가 길어지고 더 많은 수의 데이터를 사용할수록 많은 표현을 학습하고 데이터 수의 따른 성능의 변화를 확인할 수 있었다. 또한, 텍스트 데이터 실험의 경우 앞서 언급했듯이 중복되는 데이터들이 존재한다. 중복을 제거한 이후 샘플링을 진행하였고 이는 4.3에 설명되어있다. 중복되는 내용은 Table 2에서 확인할 수 있다.

Table 2. Pitch Voice Duration and Text Data Redundancy

Label	Filename	Text	Data Length
Happy	PSB_00001.json	조선인 위안부들을 구하러 한국군이 도착했습니다.	4.07
	PSB_00002.json	찾아온 것만으로 기쁜 일이라 생각했습니다.	3.26
	PSB_00003.json	아빠에게 그렇게 친절하게 대해주셔서 감사드립니다.	3.83
	LYT_00001.json	조선인 위안부들을 구하러 한국군이 도착했습니다.	3.63
	LYT_00002.json	찾아온 것만으로 기쁜 일이라 생각했습니다.	3.13
	LYT_00003.json	아빠에게 그렇게 친절하게 대해주셔서 감사드립니다.	4.23
Angry	KDD_01462.json	화를 내잖아요?	2.27
	KDD_01463.json	세 사람은 억울함에 날뛰며 자신의 결백을 증명하였다.	4.52
	KDD_01464.json	덕분에 과태료를 물었어.	3.37
	KIH_01462.json	화를 내잖아요?	1.99
	KIH_01463.json	세 사람은 억울함에 날뛰며 자신의 결백을 증명하였다.	5.02
	KIH_01464.json	덕분에 과태료를 물었어.	2.32
Neutrality	CHY_01360.json	목소리를 크게 하여 전진한다.	2.90
	CHY_01361.json	두 사람은 성의를 다해 기사를 대접하였습니다.	4.43
	CHY_01362.json	여느 때 같으면 며느리는 시아버지가 두려워 방 안에 있었다.	5.57
	LJH_01360.json	목소리를 크게 하여 전진한다.	2.62
	LJH_01361.json	두 사람은 성의를 다해 기사를 대접하였습니다.	3.99
	LJH_01362.json	여느 때 같으면 며느리는 시아버지가 두려워 방 안에 있었다.	5.00

Table 3. Accuracy Results Using Only Speech Data with Random, Short, and Long Data Sampling

Num. of Data	Random	Short	Long
5000	79.20	78.49	82.90
10000	86.32	85.89	87.20
15000	89.33	88.93	89.51
20000	92.34	89.82	92.41
25000	92.06	91.31	92.46
30000	92.79	92.25	93.70

4.2 음성 데이터 분류

[1]에서는 wav2vec 2.0 모델을 사용하여 감정 분류를 진행하였으며, 감정 분류의 성능을 향상시키기 위한 데이터 샘플링 기법을 제안하였다. 제안된 샘플링 방법에는 두 가지 접근이 포함되었다. 하나는 데이터의 양을 제한하는 방법이고, 다른 하나는 데이터의 길이가 긴 음성 데이터부터 사용하여 실험을 진행하는 방법이다.

실험 결과, 데이터의 양이 많아질수록, 그리고 데이터의 길이가 긴 음성 데이터를 사용할수록 성능이 좋아지는 것으로 나타났다. 이러한 결과는 Table 3에 자세히 나타나 있다. 표 3은 각각 데이터를 무작위, 길이가 짧은 순, 길이가 긴 순으로 데이터를 샘플링한 결과이다. 이는 음성 데이터의 양과 길이가 감정 분류 성능에 큰 영향을 미칠 수 있음을 보여주며, 효과적인 데이터 샘플링 전략이 음성 감정 분류 성능 향상에 기여할 수 있음을 보여준다. 이러한 실험 결과는 우리 연구에서 진행된 멀티모달 학습과의 비교 분석에 중요한 기준점을 제공한다.

4.3 텍스트 데이터 분류

멀티모달 학습을 적용하기 전, 텍스트 데이터만을 사용하여 감정을 분류할 때의 성능을 확인하기 위한 실험을 진행하였다. 이 실험의 목적은 텍스트 데이터만을 활용했을 때의 성능을 확인하고, 텍스트 데이터에서 학습할 표현이 많음을 보여준다. 이후 멀티모달 학습을 적용했을 때의 성능 향상을 비교 분석하는 실험에서 활용 가능성을 보여준다. 실험에 사용된 데이터는 중복을 제거한 총 15,767개의 데이터 세트이다. 중복 데이터를 제거하는 이유는 중복 데이터는 감정 분류 성능 평가의 방해 요소이기 때문에 제거하였다. 데이터 세트는 [1]에서 수행된 실험 방법을 따라 각 감정 카테고리별로 특정 개수만큼 데이터를 샘플링하여 사용하였다. 표시된 데이터 수에 7배의 데이터가 학습 데이터이다. 예를 들어 50개이면 감정별로 50개를 샘플링하여 사용했으며, 총 350개가 학습 데이터이다. all의 경우 감정별로 샘플링 없이 모든 데이터 15,767개의 데이터를 사용한 것이다.

Table 4는 텍스트만으로 감정을 분류한 결과를 정리한 것이다. 이 실험은 4.2에서 확인한 음성 데이터만을 사용한 감정 분류 실험과 비교하여, 훨씬 적은 양의 데이터를 사용하지만 7가지 감정을 효과적으로 분류하는 것을 확인하였다. 이는 텍스트 데이터가 음성 데이터에서는 학습하지 못하는 특징을 내포하고 감정을 분류하는 작업에서 사용할 가치가 충분하다는 사실을 확인하였다. 이를 통해 텍스트 데이터를 같이 사용하면 음성만을 가지고 학습을 진행할 경우보다 성능이 높아질 거라고 예상한다. 이는 4.5의 실험을 통해 확인할 수 있다.

Table 4. Accuracy Results of Text Classification Using KcELECTRA Model Based on Varying Training Data Quantities

Num. of Data	Acc(%)	Num. of Data	Acc(%)
50	72.04	500	83.93
100	75.04	1000	84.20
150	77.46	1500	84.61
200	78.92	2000	89.27
300	80.47	all	94.53

4.4 텍스트 처리 모델에 따른 실험

이번에는 음성 데이터 모델은 wav2vec 2.0으로 고정하고 텍스트 데이터를 처리하는 모델만을 변경하여 실험을 진행하였다. 이 실험의 목적은 다양한 텍스트 처리 모델들의 성능을 비교 분석하는 것이었다. 이를 통해 어떤 모델을 활용하여야 더 좋은 특징을 추출하는지 확인할 수 있다. 구체적으로, KcELECTRA[2], KoBERT[3], KLUE-bert[4], KLUE-RoBERTa[5][6], KoELECTRA[7]와 같은 여러 모델을 사용하여 각각의 성능을 평가하였다. 모든 모델은 한국어 데이터로 사전학습이 된 모델이다. KcELECTRA의 경우는 네이버 뉴스의 2019.01.01. ~ 2021.03.09. 사이에 작성된 댓글과 대댓글들을 수집하여 17.3GB의 데이터로 학습시켰다. KoELECTRA는 뉴스, 위키, 나무위키 등과 20GB의 모두의 말뭉치의 데이터를 합쳐 34GB의 데이터로 학습되었다. KoBERT의 경우 한국어 위키 데이터를 토대로 학습하였다. KLUE-bert와 KLUE-RoBERTa의 경우는 KLUE[16] 데이터 세트로 학습되었다.

Table 5를 보면 KcELECTR 모델의 성능이 제일 뛰어난 것을 확인하였다. 같은 ELECTRA를 학습시킨 KoELECTRA와 KcELECTRA를 비교할 때, KcELECTRA가 더 높은 성능을 보이는 이유 중 하나는 학습에 사용된 데이터 세트의 차이가 있기에 발생한다. KoELECTRA는 뉴스, 위키, 나무위키, 신문 등과 같은 다양한 공식적인 텍스트 소스로 학습되었다. KcELECTRA는 뉴스 댓글이나 대댓글과 같은 소셜 미디어 텍스트로 학습되었다. 이러한 데이터는 일상적인 대화와 비공식적인 언어 사용을 반영하며, 소셜 미디어 특유의 언어적 특성을 포함한다. 짧은 데이터와 구어체를 더 잘 이해한다. 실험 결과에 따라 KcELECTRA 모델을 텍스트 처리 모델로 사용하여 다음 실험을 진행하였다.

2) https://github.com/Beomi/KcELECTRA
3) https://github.com/SKTBrain/KoBERT
4) https://huggingface.co/klue/bert-base
5) https://huggingface.co/klue/roberta-large
6) https://huggingface.co/klue/roberta-small
7) https://github.com/monologg/KoELECTRA

Table 5. Performance Comparison of Different Text Processing Models - KcELECTRA, KoBERT, KLUE-bert, KLUE-RoBERTa, KoELECTRA

Model	Acc(%)
KoBERT	92.57
KoELECTRA	94.54
KLUE-bert	95.44
KLUE-roberta-small	95.49
KLUE-roberta-large	96.41
KcELECTRA	96.57

4.5 멀티모달 학습

멀티모달 학습을 적용하여 실험을 진행하였다. 실험의 데이터 수는 train 데이터 362,570개, test 데이터 45,396개, validation 데이터 45,399개 데이터에서 감정별로 샘플링하였다. 이 실험의 목적은 [1]에서 사용된 음성 데이터 기반의 샘플링 방법과 본 논문에서 진행된 텍스트 데이터 기반의 분류를 동시에 활용하는 것이다. 이를 위해 wav2vec 2.0과 KcELECTRA 모델을 사용하여 각각 음성과 텍스트 데이터로부터 특징 벡터를 추출하였다.

추출된 음성과 텍스트의 특징 벡터는 concatenate 연산을 통해 하나의 합쳐진 특징 벡터로 결합한다. 이렇게 결합한 특징 벡터는 최종적으로 FC layer를 통과하여 감정을 분류하는데 사용되었다. 이 멀티모달은 음성과 텍스트 데이터의 상호 보완적인 정보를 활용하여 감정 분류의 정확도를 높이는 것을 목표로 한다.

Table 6을 보면 멀티모달 학습을 활용하여 실험한 결과를 정리해둔 표이다. 이 실험은 멀티모달 학습을 통해 음성 데이터와 텍스트 데이터를 결합한 방식으로 진행되었다. 멀티모달 학습을 적용한 결과, 음성 데이터만을 사용하여 실험한 4.2의 결과와 비교하여 성능이 향상되었다. 이는 멀티모달 학습이 감정 분류 작업에서 음성 데이터만을 사용하는 것보다 더 효과적임을 보여준다.

음성 데이터만을 사용한 실험과 동일하게 데이터의 길이순으로 데이터를 샘플링하고 실험을 진행하였을 때 모든 경우에서 짧은 데이터가 성능이 더 낮은 것을 확인할 수 있었다. 이는 텍스트 데이터는 짧은 순서대로 데이터를 사용하지 않은 이유와 이어진다. 데이터를 짧은 순서로 사용하면 0.78, 0.89 같이 음성 데이터가 엄청 짧은 발화이며, 텍스트에서도 '뭐?' 나 '아니' 같은 엄청 짧은 텍스트 데이터가 사용되어 학습이 잘 이루어지지 않아 성능이 낮아진다. 음성과 텍스트 데이터 모두 길이가 긴 데이터를 사용하면 더 많은 표현을 학습하여 성능이 오르는 것을 확인할 수 있었다.

또한, 멀티모달 학습을 사용함으로써 훨씬 적은 데이터로도 학습이 가능하였다는 점은 데이터 자원이 제한적인 상황에서도 멀티모달 학습이 효과적으로 적용될 수 있음을 의미한다. 텍스트 데이터만을 사용했을 때와 똑같은 양의 데이터를 사용하여 멀티모달 실험을 진행한 결과, 멀티모달 학습이 훨씬 높은 성능을 보였다. 이는 멀티모달 학습이 텍스트 데이터만을 사용하는 것보다 감정 분류 작업에 더 유리함을 보여준다.

Table 6. Comparative Analysis of Multi-modal Accuracy Results: Single-Modal Speech and Text Data Versus Multi-modal Approach, Categorized by Data Length (Short and Long)

Data Type	Num. of Data	Short	Long
Speech only (best)	30000	92.25	93.70
Text only (best)	15767(all)	-	94.53
Multi-modal	50	49.13	60.53
	100	56.35	82.61
	150	68.71	89.01
	200	72.23	89.86
	250	82.70	90.31
	300	83.61	94.68
	500	84.84	96.57

추가로, [1]과의 비교를 위해 감정별 5,000개의 데이터를 사용하여 실험을 수행하였을 때, 결과는 96.93%의 정확도를 보였다. 이는 500개의 데이터로 학습하였을 때와 큰 성능 차이가 없이 정확도가 수렴하는 것을 확인할 수 있었다. 이는 멀티모달 학습이 적은 데이터로도 높은 성능을 달성할 수 있음을 나타낸다. 이러한 결과는 멀티모달 학습이 감정 분류 작업에서 음성과 텍스트 데이터의 통합적인 활용을 통해 성능을 향상시킬 수 있는 효과적인 방법임을 입증하며, 데이터의 양적 제약을 극복하고 효율적인 학습이 가능하다는 것을 보여준다.

5. 결 론

실험을 통해 얻은 결과는 멀티모달 학습이 감정 분류에 효과적임을 보여준다. 4.2에서 수행된 음성 데이터만을 사용한 실험과 4.5에서 진행된 실험을 비교 분석한 결과, 멀티모달 학습을 적용하여 감정 분류를 진행했을 때 정확도가 더 높게 나타났다. 이는 음성과 텍스트 데이터를 결합하는 멀티모달 학습이 단일 데이터 방법에 비해 감정 분류 작업에서 더 우수한 성능을 제공함을 의미한다.

결과를 통해 길이가 긴 데이터를 사용할수록 성능이 높아지는 것을 확인할 수 있었다. 이를 통해 데이터를 효율적으로 활용하는 샘플링 방법이 성능에 영향을 주는 것을 확인하였다.

또한, 멀티모달 학습은 더 적은 데이터로도 높은 정확도를 달성할 수 있음을 보여주었다. 이는 데이터 자원이 제한적인 상황에서도 멀티모달 학습이 효과적으로 적용될 수 있음을 보여주며, 특히 데이터 수집에 어려움을 겪는 음성 인식 및 자연어 처리 분야에서 의미가 있다.

멀티모달 학습을 적용할 때 텍스트 처리 모델의 선택이 매우 중요하다. 실험을 통해 다양한 모델들의 성능을 비교한 결과, 발전된 모델(ELECTRA, RoBERTa)이 기존 모델(BERT)보다 더 높은 정확도를 달성함을 확인할 수 있었다. 같은 모델이더라도 학습 데이터에 따라 성능에 변화가 있음을 알 수 있었

다. 따라서 실험에 적합한 데이터 세트로 학습된 모델을 선택하는 것이 멀티모달 학습에 성공적인 요소이다.

이번 연구는 멀티모달 학습이 감정 분류 작업에서 어떻게 성능을 개선할 수 있는지에 대한 실질적인 이해를 제공한다.

References

[1] M. Shin and Y. Shin, "Data sampling strategy for Korean speech emotion classification using wav2vec2.0," in *Proceedings of the Annual Conference of Korea Information Processing Society Conference (KIPS) 2023*, Vol.30, No.2, pp.493-494, 2023. [Online]. Available: https://kiss.kstudy.com/Detail/Ar?key=4059338

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, Vol.33, pp.12449-12460, 2020.

[3] S., Schneider, A., Baevski, R., Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[4] A., Vaswani et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, Vol.30, 2017.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[6] Y. Liu, et al., "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[7] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555*, 2020.

[8] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," In: *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. pp.112-118, 2018.

[9] X, Zhang, M, Wang, and X. Guo, "Multi-modal emotion recognition based on deep learning in speech, video and text," *2020 IEEE 5th International Conference on Signal and Image Processing (ICSIP), Signal and Image Processing (ICSIP)*, pp.328-333, 2020.

[10] J. Agarkhed, "Machine learning based integrated audio and text modalities for enhanced emotional analysis," In *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, pp.989-993, 2023.

[11] S. S. Hosseini, M. R. Yamaghani, and S. Poorzaker Arabani, "Multimodal modelling of human emotion using sound, image and text fusion," *Signal, Image and Video Processing*, Vol.18, No.1, pp.71-79, 2024.

[12] H. Koh, S. Joo, and K. Jung, "Reflecting dialogue and pretrained information into multi modal emotion recognition: focusing on text and audio," in *The Korean Institute of Information Scientists and Engineers*, pp.2136-2138, 2023, [Online]. Available: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11488656

[13] Y.-J. Kim, K. Roh, and D. Chae, "Feature-based Emotion Recognition Model Using Multimodal Data," in *The Korean Institute of Information Scientists and Engineers*, pp. 2169-2171. 2023, [Online]. Available: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11488667

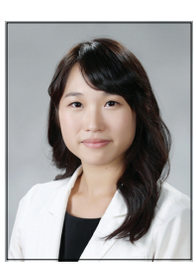
[14] T. Yoon, S. Lee, H. Lee, H. Jin, and M. Song, "CoKoME: Context modeling for Korean multimodal emotion recognition in conversation," in *The Korean Institute of Information Scientists and Engineers*, pp.2100-2102, 2023, [Online]. Available: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11488644

[15] J. Lee, J. Bae, and S. Cho, "Multi-modal emotion recognition in Korean conversation via Contextualized GNN," in *The Korean Institute of Information Scientists and Engineers*, pp.2094-2096. 2023, [Online]. Available: https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11488642

[16] S. Park, et al., "Klue: Korean language understanding evaluation," *arXiv preprint arXiv:2105.09680*, 2021.

신 미 르

https://orcid.org/0009-0004-4430-0491
e-mail : tlsalfm820@inu.ac.kr
2024년 인천대학교 컴퓨터공학부(학사)
2024년 ~ 현 재 인천대학교 컴퓨터공학과 석・박사통합과정
관심분야 : 자연어 처리, 음성 인식, 음성 언어 이해, 인공지능



신 유 현

https://orcid.org/0000-0001-7013-9057
e-mail : yhshin@inu.ac.kr
2019년 서울대학교 컴퓨터공학부(박사)
2020년 ~ 현 재 인천대학교 컴퓨터공학부 부교수
관심분야 : 자연어 처리, 음성 언어 이해, 인공지능