# 멀티 모달 데이터를 활용한 결정 융합 기반 감정 분류 모델

이서연 [O1] 김원중 [2]
[1] 서강대학교 커뮤니케이션학과
[2] 홍익대학교 글로벌경영학과
0327seoyeon@gmail.com, kwj102501@naver.com

# An emotion classification model based on decision fusion using multi-modal data

Seoyeon Lee[O1] Wonjoong Kim[2]
[1]Department of Communication, Sogang University
[2]Department of Global Business Administration, Hongik University

**요 약**

본 연구에서는 인공 신경망을 통해서 생체 신호, 텍스트, 음성 데이터를 복합적으로 활용하여 기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔의 7 가지 감정을 판별하는 분류 모델을 생성하는 것을 목표로 한다. 해당 연구에서 사용한 데이터인 KEMDy19[1] 는 남녀 각각 20 명의 성우가 2 명씩 한 조를 이루어 감정 상황극을 연기하고, 그 과정에서 성우의 발화 음성, 발화 텍스트, 생체 신호를 수집한 것이다. 각각의 멀티 모달 데이터의 특성에 맞는 딥러닝 모델을 활용하여 감정 분류 모델을 생성하였다. 이후 생성된 모델들의 성능을 더욱 향상하기 위해 앙상블 모델 중 soft voting 의 개념을 차용하여 실험을 진행하였고, 그 결과 단일 모델로 감정 분류를 했을 때와 비교하여 약 12%의 향상된 정확도를 보였다.

## 1. 서 론

최근에 인공지능 기술의 발달로 인간-컴퓨터 상호작용(HCI) 분야에 새로운 커뮤니케이션 패러다임이 펼쳐지고 있다. 그러나 오늘날까지 인간이 인공지능과 소통하는 과정에서 느끼는 이질감은 해결되지 못한 상황이다.

사람은 기본적으로 소통하는 과정에서 언어적 메시지와 비언어적 단서(nonverbal clues)를 모두 사용하여 정보를 처리하는데, 상대방에 대한 정보가 부족한 상황에서는 비언어적 단서가 상대방에 대한 신뢰와 호감을 형성하는 데 중요한 영향을 미친다[1]. 인간과 인공지능 간의 소통에서 발생 되는 이질감은 이러한 비언어적 단서의 부재로부터 시작된다.

과거에는 HCI 분야에서 겉으로 드러나지 않는 내면적인 감정을 파악하는 것이 어려웠다. 그러나 오늘날에는 생체 신호 정보와 심리 상태의 상관관계가 크다는 것이 밝혀져, 감정 인식을 하는 데 생체 정보가 활용될 수 있다[2].

따라서 텍스트와 같은 언어적 메시지뿐만 아니라 생체 신호 등의 비언어적인 단서까지 활용한 감정 인식 기술은 진정한 인간 중심의 human to machine interface 의 구현을 지향하며, HCI 분야의 큰 발전에 기여하게 될 것이다. 더 나아가 인공지능 기술에 대한 사람들의 수용도가 향상되어, 인공지능 기술이 시장에 정착하는 데 가속화될 것으로 전망한다.

기존의 감정 인식 기술 분야에서는 유니 모달 데이터를 활용한 기술이 주류였다. 그러나 감정은 단일 요소로 결정되기보다는 복합적인 내·외부적 요인으로부터 영향을 받기 때문에 유니 모달 데이터만을 활용하는 것은 감정 인식 정확도가 다소 불안정하다. 따라서 본 연구에서는 단일 신호에만 의존하지 않고 다양한 형태의 멀티 모달 신호를 사용하기로 한다.

## 2. 감정 분류 기술 구현 방안

### 2.1 결정 융합 기반의 멀티 모달 모델

멀티 모달 데이터를 활용한 감정 분류 방법을 크게 2 가지로 구분하였다.

첫 번째는 특징 차원에서의 융합 방법으로 각각의 멀티 모달 데이터들의 차원을 일치시킨 다음, 정규화를 거쳐 특징값들을 융합하고 모델 학습에 적용하는 것이다.

두 번째로는 결정 값 차원에서의 융합 방법으로 멀티 모달 데이터를 각각의 분류 모델에 적용한 다음, 그 결괏값을 융합하는 방법이다[3].

특징 기반 융합의 경우 연산량이 적다는 장점이 있지만 아래와 같은 문제를 야기시킨다.

1) **각 모달리티 데이터가 갖는 고유의 특성 손실**
각 모달리티 데이터의 특징값을 단순히 정규화해 결합할 경우 각각의 데이터가 표현하는 의미 및 특성이 손실된다.

2) **자료형 구조가 상이한 데이터를 융합하는 경우 자료형의 특이성이 상실**

[1] https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR

본 연구에서 사용하는 음성 및 생체 신호 데이터의 자료형은 수치형 데이터에 속하고 텍스트 데이터의 경우 범주형 데이터로 속한다. 이들을 융합하게 되면 각각의 데이터 구조가 갖는 특이성이 상실된다.

3) **융합되는 데이터 간의 특징값 개수의 격차가 심한 경우 예측 결과의 편향이 발생**

본 연구에서 사용되는 멀티 모달 데이터들은 한 데이터가 갖는 특징값이 생체 신호 데이터의 경우 3 개, 텍스트 데이터는 94 개, 음성 데이터는 46,000 개로 구성된다. 이들을 융합하여 하나의 새로운 특징값을 만들어 모델에 적용할 경우 그 예측 결과는 특징값이 상당히 많은 음성 데이터의 정보에 의존한 결과가 나오게 된다.

이러한 점을 고려하였을 때 본 연구에서는 결정 융합을 기반으로 한 분류 모델 개발을 목표로 진행한다.

### 2.2 연구 모델 구조

감정 분류 모델의 구성은 크게 3 가지 독립변수의 개별 분류 모델과 이들을 통합하는 앙상블 모델로 이루어진다.

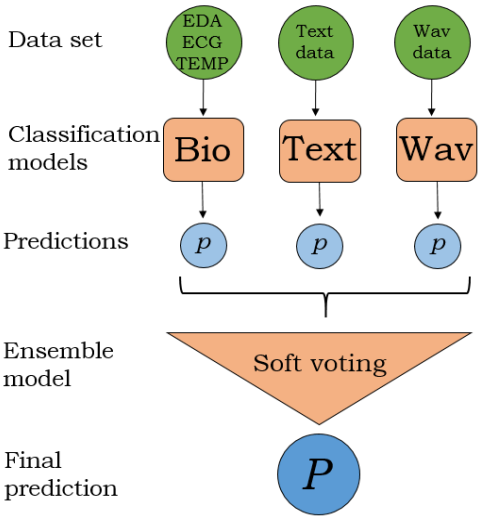


그림 1 감정 분류 모델 프로세스

그림 1 과 같이 먼저 생체 신호, 텍스트, 음성 데이터를 기반으로 한 독립적인 3 개의 감정 분류 모델에서 예측값을 도출한다. 이후 각 모델에서 나온 예측값을 앙상블 모델에 활용하여 최종값을 결정하는 구조다.

### 2.3 불균형 데이터(imbalanced data)의 문제 해결

KEMDy19 데이터셋에서 다소 심각한 수준의 클래스 불균형 문제가 발견되었다. 최대 개수의 데이터를 갖는 클래스인 Neutral 과 최소 개수의 데이터를 갖는 클래스인 Surprise 사이에서 약 17.3 배의 데이터 개수 차이가 나타났다.

본 연구에서는 부족한 데이터를 SMOTE(Synthetic Minority Oversampling Technique) 처리를 통해 증강하여 클래스별 편향이 발생하지 않도록 한다.

## 3. 독립 변수별 분류 모델 생성

### 3.1 생체 신호 기반 감정 분류 모델

본 연구에서 사용되는 생체 신호 정보의 데이터셋은 EDA, ECG, TEMP 의 3 가지 데이터를 하나의 벡터로 조합하여 구성되어있다. 이를 아래 그림 2 와 같은 DNN(Deep Neural Network) 모델에 적용하여 학습시킨다.

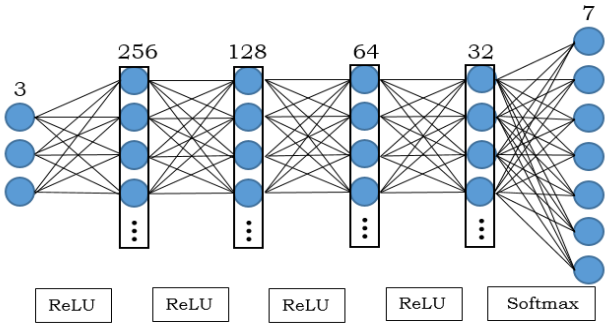


그림 2 생체 신호 데이터를 활용한 DNN 모델의 구조

### 3.2 텍스트 정보 기반 감정 분류 모델

그림 3 과 같이 데이터 전처리 과정에서 텍스트 데이터의 형태소 분석을 위해 Mecab 라이브러리를 활용한다. 토큰화된 텍스트에서 Stop words 를 통해 '가', '이', '도', '는' 등의 불용어를 설정하여 제거한다. 이후 텍스트 시퀀스로 된 데이터를 정수 시퀀스로 변환하고 데이터의 최대 길이에 맞춰 패딩 작업을 거친다.

본 연구에서 사용될 텍스트 데이터의 양이 다소 부족하다는 점을 고려 하였을 때 RNN 계열 중 LSTM 과 비교하여 상대적으로 매개 변수의 양이 적은 GRU 모델을 사용하기로 한다.

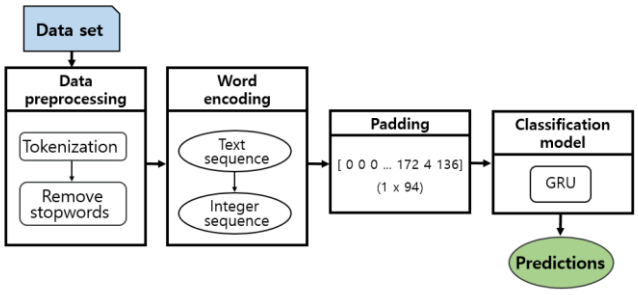


그림 3 텍스트 데이터를 활용한 감정 분류 모델 프로세스

### 3.3 음성 데이터 기반 감정 분류 모델

먼저 전처리 과정에서 librosa 라이브러리를 활용하여 MFCC(Mel-Frequency Cepstral Coefficient)를 추출하고

특징값을 얻는다. 이후 이상치를 제거한 후 데이터의 평균 길이에 맞춰 패딩 작업을 거친다.

본 논문에서는 그림 4 와 같이 MFCC 값을 기반으로 한 감정 분류 모델로 CNN(Convolutional Neural Network)을 활용한다.

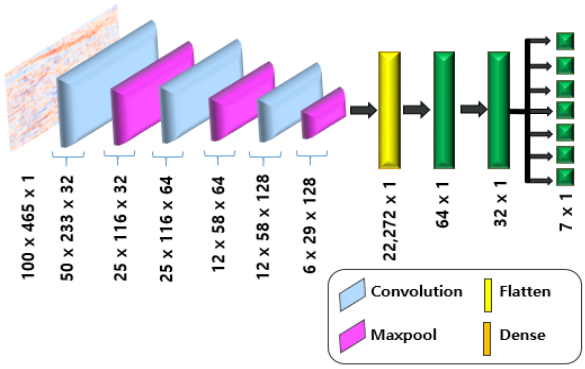


그림 4 음성 데이터를 활용한 CNN 모델 구조

## 4. 앙상블

결정 융합의 방법으로 앙상블 기법 중 soft voting 방법을 활용하고 가중치를 동일하게 설정한다.

soft voting 은 각 모델이 예측한 probability 를 클래스별로 합산하여 가장 높은 클래스를 최종 결과로 결정한다. 아래의 식 (1)과 같이 계산한다.

$$\hat{y} = \arg\max_{i} \sum_{j=1}^{m} p_{ij} \quad (1)$$

투표에 참여하는 모델의 수를 m, 클래스가 7 개일 때 각 모델에서 도출되는 probability 를 [p01, p11 … p71], [p02, p12 … p72] … [p0m, p1m … p7m]이라고 표현한다. p01 + p02 + … +p0m 과 같이 클래스별로 동일한 위치에 있는 probability 를 합산하여 최댓값을 찾는다.

## 5. 성능평가

표 1 모델별 성능 평가 비교

| | | Recall | Precision | F1 score |
|---|---|---|---|---|
| Models | Bio (DNN model) | 0.556 | 0.579 | 0.537 |
| | Text (GRU model) | 0.779 | 0.794 | 0.771 |
| | Wav (CNN model) | 0.738 | 0.792 | 0.751 |
| | Ensemble model | 0.805 | 0.844 | 0.806 |

독립 변수별로 학습한 모델들의 성능과 이들을 하나로 통합한 앙상블 모델의 성능을 비교하였다. F1 score 를 기준으로 각 모델의 성능을 평가한 결과, 표 1 과 같은 성능을 보였다.

각 독립 변수 기반 모델과 앙상블 모델 사이의 성능 향상률을 살펴보면 앙상블 모델을 사용 시, 생체 신호 기반 모델에서는 26.9%, 텍스트 기반 모델에서는 3.5%, 음성 기반 모델에서는 5.5%의 성능 향상을 기록하였다.

따라서 본 연구에서는 유니 모달 데이터 기반의 모델에서 앙상블을 통한 멀티 모달 데이터 기반의 모델을 구축할 경우, 평균적으로 약 12%의 성능 향상이 있다는 것을 알 수 있다.

## 6. 결론 및 향후 과제

본 논문에서는 단일 모달 데이터만을 활용한 감정 분류 모델과 앙상블 기법을 통해 다양한 멀티 모달 데이터를 활용하여 하나로 통합한 감정 분류 모델 사이의 성능을 비교해보았다. 생체 신호 데이터 기반 모델 등 성능이 다소 아쉬운 모델도 다른 독립 변수 기반의 모델과 통합하는 경우 성능이 향상된다는 것을 알 수 있다.

본 연구 주제와 관련된 향후 연구 과제는 결정 융합이 아닌 특징 융합으로 멀티 모달 데이터를 통합하여, 이를 앙상블 모델 중 Bagging 기법을 활용해서 감정 분류하는 모델을 구축해 보는 것이다. 독립 변수가 A, B, C 의 3 가지로 구성된다고 가정했을 때, 융합할 변수들의 자료형을 고려하면서 A 와 B, B 와 C, A 와 C 등 모든 경우의 수로 조합하여 특징 융합의 단점을 보완하고자 한다. 각 조합의 독립 변수를 동일한 종류의 모델에 적용하여 도출한 예측값을 최종적으로 voting 한다.

이렇게 최적의 성능을 보이는 앙상블 모델을 구축하게 되면 감정 분류 모델을 떠나, 다양한 멀티 모달 데이터와 모델을 자유롭게 조합할 수 있게 된다. 향후 인공지능이 복합적이고 정교한 인간의 세계를 더욱 고차원적으로 이해하게 되어 인간과 인공지능이 공존하는 날이 앞당겨질 것이다.

## 참고 문헌

[1] 송근혜, 이승민, “기술진화로 인한 인간과 기계의 사회적 관계 연구”, 한국기술혁신학회 학술대회, pp. 483-496, 2017.

[2] 송병철, 김대하, 최동윤, 이민규, “감정 인식 기술 동향”, ICT 신기술, 18-29, 2018.

[3] 고광은, 심귀보. “멀티 모달 감정인식 시스템 기반 상황인식 서비스 추론 기술 개발”, 한국지능시스템학회 학술발표 논문집, Vol. 19, No. 1, pp. 34-39, 2009.