멀티 모달 기반 감정 인식 모델에 대한 연구

강찬혁, 한경식
아주대학교
{rkdcksgur, kyungsikhan}@ajou.ac.kr

A Study for Emotion Recognition Model based on Multi-Modality

Chanhyuk Kang, Kyungsik Han
Ajou University

요 약

인공지능 기술이 발전하고 고도화되면서 각각의 모달리티에서 얻은 데이터를 통합적으로 해석하는 것이 중요해지고 있다. 이에 따라 본 연구에서는 텍스트, 이미지, 음성 3가지의 모달리티를 통합한 멀티 모달리티로 드라마의 감정 인식 연구를 진행하였다. 3가지 감정(긍정, 중립, 부정)으로 분류한 데이터를 바탕으로 머신러닝 모델을 구축하였다. 단일 모달리티에서의 결과를 정성적으로 분석하고 문장과 대화 간의 관계와 중요도를 파악하는 멀티 모달리티 모델을 구축하였다. 그 결과, 멀티 모달리티에서의 감정 인식 성능이 가장 높은 성능을 보였고 사람이 감정을 분류하는 방식보다도 성능이 향상됨을 확인하였다.

Ⅰ. 서 론

최근 인공지능 기술이 발전하고 고도화됨에 따라 인간과 컴퓨터 사이의 인터랙션의 중요성이 강조되고 인간의 감정을 다양한 인터페이스를 통해 파악하고 분석하는 멀티 모달리티[1] 서비스·시스템 등의 필요성이 높아질 것으로 예상된다. 따라서 각각의 모달리티에서 얻은 데이터 사이의 관계를 통합적으로 해석할 수 있는 것이 중요하다[2].

본 연구에서는 텍스트, 이미지, 음성 3가지의 모달리티를 통해 데이터를 수집하고 가공한다. 다양한 입력 모달리티에서 얻은 데이터를 바탕으로 각각의 모달리티 사이의 관계를 통합적으로 해석할 수 있는 모델을 만들고 연구한다. 그리고 각각의 모달리티에서의 dialogue와 utterance 사이의 감정 관계를 분석해 인간의 말에 담긴 감정을 분석할 수 있는 더 나은 방법과 사용자 인터페이스와 경험을 제공할 수 있는 방법을 제안한다.

Ⅱ. 본론

1. 데이터 수집

텍스트, 이미지, 음성 총 3가지 모달리티에서 감정을 추출하기 위한 데이터셋은 미국 NBC에서 1994년부터 2004년까지 방송했던 시트콤 <프렌즈>에서 한 문장씩 분리된 11384개의 데이터셋을 사용한다. anger, happiness, sadness, fear, surprise, neutral, disgust 7개의 감정으로 구성되어 있고 이번 연구에서는 neutral, happiness, sadness, anger 4가지 감정을 사용했다. (neutral 6073개, happiness 2675개, sadness 953개, surprise 806개, disgust 403개, anger 374개, fear 99개) 그리고 이전의 연구에서 문장 별로 나누어진 <프렌즈> 텍스트와 이미지 데이터에서 얻은 감정 레이블을 함께 사용한다.

음성 모달리티에서 감정을 얻기 위한 머신러닝 학습 용도로는 IEMOCAP(Interactive emotional dyadic motion capture database) 데이터셋을 사용한다. happy, sad, neutral, angry, frustrated, excited 6개의 감정으로 구성되어 있고 음성 데이터는 5606개이다. 이번 연구에서는 happy, sad, neutral, angry 4가지의 감정을 사용했다. (frustrated 1393개, neutral 1295개, angry 890개, sad 799개, excited 725개, happy 504개)

2. 데이터 전처리

음성에서 감정을 얻기 위해서는 드라마 <프렌즈> 영상으로부터 한 문장으로 이루어진 음성 데이터를 뽑아내야 하고 음성의 고유한 특징을 수치로 나타내는 특징을 바탕으로 머신러닝 모델을 설계해야 한다.

영상으로부터 음성 데이터를 추출하기 위한 라이브러리로 moviepy.editor를 사용했다. 그리고 에피소드 별 음성에서 배우들의 대사 별 문장 음성 데이터를 얻기 위한 라이브러리로 pydub과 ffmepg[5]을 사용했다. 분리된 문장 별 음성 데이터에서 librosa, subprocess, pyAudioAnalysis를 사용해 음성의 고유한 특징인 MFCC(Mel-Frequency Cepstral Coefficient), chroma, 에너지 등의 값을 얻었다[3][4]. 음성 모달리티에서 감정 레이블을 얻은 후 happiness는 positive, sad와 anger는 negative로 변경했다.

단일 모달리티와 멀티 모달리티의 감정 분류 성능을 비교하기 위해 scikit-learn[6]에서 제공하는 TF-IDF(Term Frequency Inverse Document Frequency) 방식으로 특징을 추출했다. 각각의 모달리티의 성능을 비교하기 위해 텍스트, 이미지, 멀티 모달리티 감정 레이블도 마찬가지로 happiness는 positive, sad와 anger는 negative로 변경해 평가했다.

3. 모델 구축

감정 인식 모델 구축을 위해 개발 언어로는 파이썬을 사용했고 머신러닝 학습을 위해 scikit-learn[6]에서 제공하는 XGBoost와 RandomForest 알고리즘을 사용했다. 멀티 모달리티 감정 인식 과정은 3가지의 단일 모달리티에서의 감정 분류, 정성 평가 후 멀티 모달리티 감정 분류, 텍스트 임베딩, 성능 평가 순으로 진행된다.

먼저 음성 모달리티에서의 감정 분류를 위한 머신러닝 학습은 Training set으로 IEMOCAP 데이터셋, Test set으로 Friends 데이터셋을 사용했다. 감정 인식을 위한 알고리즘은 XGBoost를 사용했고 사용된 특징은 MFCC를 포함한 76개이다.

멀티 모달리티 감정 분류는 이전의 연구에서 진행된 텍스트와 이미지 모달리티 감정 레이블과 함께 음성 모달리티 감정 레이블을 정성적으로 분석해 결정했다. 하나의 문장(Utterance)과 Dialogue의 관계와 3개의 단일

모달리티에서의 감정 레이블 결과를 아래의 표로 정리했다.

그림1. Utterance #1, #2, #3의 장면

표1. Utterance #1 모달리티 별 Sentiment

Utterance	I did.		
Modality	Text	Image	Audio
Sentiment	neutral	positive	neutral

표2. Utterance #2 모달리티 별 Sentiment

Utterance	...I accept all those flaws, why can't you accept me for this?		
Modality	Text	Image	Audio
Sentiment	negative	negative	positive

표3. Utterance #3 모달리티 별 Sentiment

Utterance	Oh, he needed some time to grieve. That oughta do it.		
Modality	Text	Image	Audio
Sentiment	neutral	positive	positive

Dialogue #1은 남자 배우가 어릴 적 여자 배우를 좋아했다고 말해주는 상황이다. 얼굴의 미소를 이미지 모달리티에서는 긍정으로 분류했지만 문맥상 중립이 맞다고 판단했다.

Dialogue #2는 친구들이 남자 배우에게 담배 피지 말라고 잔소리를 하는 상황이다. 억양이 높아지고 말꼬리가 올라가는 것을 음성 모달리티에서는 긍정으로 분류했지만 문맥상 부정이 맞다고 판단했다.

Dialogue #3는 남자 배우가 소개팅에서 만난 여자를 맘에 들지 않아 했는데 더 이상 만나지 않는다는 것을 알게 되는 상황이다. 반어법을 사용하고 있는 상황인데 텍스트 모달리티에서는 중립으로 분류했다. 하지만 문맥상 긍정이 맞다고 판단했다.

위와 같이 정성적인 분석을 통해 3가지 단일 모달리티에서의 감정 레이블 중 다수를 따르는 감정을 멀티 모달리티의 감정 레이블로 정했다. 3가지 단일 모달리티에서 레이블이 겹치지 않는 경우는 각 Utterance마다 감정이 일관성을 띄지 않아 하나의 감정으로 정할 수 없다고 판단하여 neutral로 최종 결정했다.

이후 멀티 모달리티에서의 최종 감정 레이블과 3가지 단일 모달리티 감정 레이블의 성능을 비교하기 위해 텍스트 임베딩을 진행했다. TF-IDF vectorization를 사용해 하나의 Uttreance를 어떤 단어가 Dialogue 내에서 얼마나 중요한 지를 나타낼 수 있는 실수 벡터 값으로 변환시켜 특징으로 사용했다. 4가지 방식에 따른 감정 인식 성능 비교를 위한 머신러닝 알고리즘은 RandomForest를 사용했다.

4. 성능

텍스트, 이미지, 음성 3가지의 단일 모달리티로 얻은 감정 레이블과 멀티 모달리티로 얻은 감정 레이블의 정확도를 평가하기 위해 Training Set 80%, Test Set 20%로 데이터를 나누어 5-Fold 교차 검증을 수행했다. 각 모달리티의 성능 결과를 아래의 표로 정리했다.

표4. 모달리티 별 감정 인식 성능

	Accuray	F1-Score	Precision	Recall
Text	0.520	0.445	0.499	0.449
Image	0.414	0.391	0.398	0.401
Audio	0.565	0.383	0.377	0.392
Multi-Modal	**0.476**	**0.468**	**0.478**	**0.476**

4가지 방식 중 멀티 모달리티 방식의 F1-Score가 가장 높은 것을 확인하였다.

Ⅲ. 결론

본 논문에서는 감정 인식을 위해 사람과 같은 방식으로 동시에 여러 모달리티로 감정을 예측하는 분류 모델을 구축했다. 단일 모달리티와 멀티 모달리티의 성능을 비교하기 위해 음성 처리, 정성 평가, 텍스트 벡터화를 진행했고 이번 연구 방식이 감정을 인식하는데 더 나은 방식이라는 것을 확인했다. 특히 텍스트 모달리티의 감정 레이블은 사람이 직접 감정을 분류한 결과이기 때문에 텍스트 모달리티보다 멀티 모달리티 방식의 성능이 높은 것은 의미가 있다.

하지만 4가지 방식 모두 정확도가 비교적 낮았는데 텍스트 모달리티에서는 사람의 주관적 요소로 인해, 이미지 모달리티에서는 화자가 동시에 빠르게 전환되거나 배경 및 주변 인물로 인해, 음성 모달리티에서는 파형 분석과 함께 언어 해석이 불가하고 반어법 등의 이유로 인해 한계점이 있다고 판단했다.

머신러닝 모델의 성능을 높이거나 이번 연구에서 진행한 정성적 평가를 일반화해 적용하거나 문맥적인 요소를 포함하도록 텍스트 벡터화를 진행한다면 정확도는 더 높아질 것이다. 향후 연구를 이러한 방향으로 진행한다면 기계도 사람과 같은 방식으로 보다 정확하게 감정을 분류할 수 있고 여러 비언어적인 모달리티에서 얻은 정보를 통합적으로 해석해 감정을 인식하고 사람의 감성과 경험을 위한 서비스에서 유의미하게 활용될 것으로 기대한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 수행 결과로 추진되었음" (2015-0-00908)

참 고 문 헌

[1] 임미정 외, "멀티모달 인터페이스 개발을 위한 휴먼-컴퓨터 인터랙션 설계". 2006. (http://www.w3.org/TR/mmi-reqs/)

[2] S. Oviatt, "User-centered modeling and evaluation of multimodal interfaces," in Proceedings of the IEEE, vol. 91, no. 9, pp. 1457-1468, Sept. 2003, doi: 10.1109/JPROC.2003.817127.

[3] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," Proc. Eurospeech, Geneva, Switzerland, pp. 125-128, 2003.

[4] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov modelbased speech emotion recognition," Proc. ICASSP, Hongkong, China, pp. 401-404, 2003.

[5] ffmpeg library: http://ffmpeg.org/

[6] Scikit-learn library: http://scikit-learn.org/