**딥러닝 기반의 방법론을 이용한 텍스트 데이터 분석**

**- BERT를 이용한 음원 가사 다중 감성(emotion) 예측**

**목차.**

**1. 연구 동기 및 목적 (Motivation and Purpose)**

**2. 데이터와 연구 방법론 (Data and Methods)**

**2.1 데이터 수집 및 전처리**

**2.2 모델링**

**3. 분석 결과(Results)와 시사점**

**3.1 EDA**

**3.2 Clustering**

**3.4 시사점**

**4. python code, 참고문헌 및 자료**

**1. 연구 동기 및 목적 (Motivation and Purpose)**

본 연구는 음원 스트리밍 서비스의 Top100 차트에 진입하는 음원들의 내재적 특성을 알아보기 위해 진행되었다. 자연어 처리 딥러닝 모델인 BERT를 활용하여 음원의 가사가 전달하는 감정을 인식 및 추출하였다. 따라서 본 논문은 정형 데이터인 음원의 오디오 특성뿐만 아니라 비정형 데이터인 가사(lyrics)의 특성도 함께 고려했다는 점에서 의의가 있다.

음악 감정 인식 및 분류는 음악 정보 검색(MIR) 분야의 하위 영역으로, 음악 추천 시스템을 정교화하기 위한 방안을 연구하는 많은 연구자들의 관심사였다(Rajendran, R. V., Pillai, A. S., & Daneshfar, F. (2022)). 심리학의 여러 감정 모델 중 James Russel의 원형감정모델(Russell’s circumplex model)이 MIR 연구에서 가장 많이 사용되어왔다(Hu, X., & Downie, J. S. (2010)). 러셀의 모델은 감정이 연속적인 다차원에 위치하는 차원 모델로, Valence(부정적-긍정적)과 Arousal(비활성-활성)의 두 가지 차원으로 이뤄져 있다. 아래는 28개의 감정 형용사를 이 두 차원으로 이뤄진 양극성 공간에 배치한 그림이다(Hu, X., & Downie, J. S. (2010)).

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

본 연구는 Russell 감정 모델에 기반한 라벨링된 데이터 확보에 어려움이 있어,음원 감정 분류 범주로 Paul Ekman(1992)의 6가지 감정(anger, disgust, fear, hapinness, sadness, surprise) 모델을 사용했다.

**2. 데이터와 연구 방법론 (Data and Methods)**

**2.1 데이터 수집 및 전처리**

* 음원데이터:   
  Spotify API 서비스를 활용하여, Spotify의 “Top Hits of 2015”, “Top Hits of 2016”, “Top Hits of 2017”, “Top Hits of 2018”, “Top Hits of 2019” 재생목록 각각에 수록된 100개의 음원과 각 음원의 정보를 수집하였다.  
  결과적으로 총 500개의 음원에 대하여 16개의 음원 정보(track\_name, artist\_name, release\_date, popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration\_ms, )를 확보하였다.

\*Spotify: 글로벌 음악 및 팟캐스트 스트리밍 플랫폼으로, 스트리밍 음악 서비스 기준 전 세계에서 가장 많은 이용자(약 3.5억명 이상)를 보유하고 있다.

* 가사데이터:   
  Spotify API 서비스에서 음원의 가사 정보를 제공하지 않아, Genius API 서비스를 활용하여 각 음원의 track\_name과 artist\_name을 기준으로 가사를 크롤링했다.  
  모델 예측을 원활하게 하기 위해, 음원의 가사를 줄바꿈(\n) 기준으로 나눈 뒤 토큰화를 진행하였다.  
  \*Genius: 가사 검색 엔진 및 음악 정보 플랫폼  
  스크린샷, 텍스트, 라인, 직사각형이(가) 표시된 사진

  자동 생성된 설명

그림 1 2015-2019년 가사 문장 기준 토큰 개수 분포

**2.2 모델링**

* Emotion English DistilRoBERTa-base:  
  Hugging Face에 업로드된, 영어 텍스트 데이터에서 감정을 분류하는 모델이다. DistilRoBERTa-base을 파인튜닝하여, Ekman(1992)의 감정 모델에 기반한 6가지 감정(anger, disgust, fear, joy, sadness, surprise)과 neutral 감정을 예측한다. 학습에 사용된 데이터셋으로는 트위터, Reddit, 학생의 자기 보고서, TV 대화의 발화에 대한 감정 레이블이 포함되어있다.  
  텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

  자동 생성된 설명  
  위에 나열된 데이터셋의 balanced-subset(감정당 2,811개, 즉 총 20,000개)으로 학습되었다. 80%는 훈련에, 20%는 평가에 사용되었고, 평가 정확도는 66%이다(무작위 확률 기준선인 1/7 = 14% 대비).

이 모델을 활용해 499개\* 곡의 30,505개 가사 문장에 대한 감정 예측을 수행하였다.  
\*Top hits of 2015에 수록된 한 곡이 연주곡으로, 가사가 없어 제외됐다.

* 모델 예측 결과:  
  Emotion English DistilRoBERTa-base을 이용해 가사 문장의 감정 예측값(pred, label)과 7가지 감정별 score값\*을 구했다.  
  \*neutral, anger, disgust, fear, joy, sadness, surprise

스크린샷, 텍스트, 흑백, 블랙이(가) 표시된 사진

자동 생성된 설명  
연도별로 가사 문장의 감정 예측값 분포를 확인한 결과 neutral이 타 감정 대비 3~12배 많았다.

스크린샷, 텍스트, 도표, 다채로움이(가) 표시된 사진

자동 생성된 설명

그림 2연도별 문장 감정 분포 (누적막대그래프)

텍스트, 도표, 그래프, 스크린샷이(가) 표시된 사진

자동 생성된 설명

그림 3연도별 가사 문장 감정 분포

문장이 아닌 전체 가사 기준으로 집계했을 때 가장 많은 빈도를 보인 label을 음원이 주요 감정으로 할당한 뒤 확인한 결과, 2015년부터 2019년까지 차례대로 86, 86, 99, 95, 91개의 곡의 주요 감정이 neutral로 예측되었다.  
 본 연구는 음원의 가사별 감정의 차이에 주목하기 위해, neutral을 제거한 뒤 분석을 진행하였다. 음원별로 가사 문장의 각 6가지 감정의 score값을 합하고, 그렇게 구한 총합 대비 가장 큰 값을 가진 감정을 음원의 주요 감정(pred\_emotion)으로 결정하였다.

|  |  |  |
| --- | --- | --- |
| Column | 설명 | 비고 |
| track\_name | 음원 이름 | key |
| artist\_name | 가수 |  |
| popularity | 유명한 정도 | 0~100 |
| danceability | 춤을 추기 적합한 정도 | 0~1 |
| energy | 활동성과 에너지 수준 | 0~1 |
| key | 음계(키) |  |
| loudness | 음량 크기 | dB(데시벨) |
| mode | 음계 모드 | 0(minor), 1(major) |
| speechiness | 말(발성)의 비중 | 0~1 |
| acousticness | 어쿠어스틱한 정도 | 0~1 |
| instrumentalness | 악기 중심인 정도 | 0~1 |
| liveness | 실확 녹음 정도 | 0~1 |
| valence | 긍정적 감성 정도 | 0~1 |
| tempo | 빠르기 정도(분당 박자) | BPM |
| duration\_ms | 음원 | ms(밀리세컨드) |
| added\_year | 발매일\_연도 |  |
| added\_month | 발매일\_월 | 1~12 |
| added\_day | 발매일\_요일 | Mon, Tue, … |
| anger | 분노 수준 | 0~1 |
| disgust | 혐오 수준 | 0~1 |
| fear | 두려움 수준 | 0~1 |
| joy | 기쁨 수준 | 0~1 |
| sadness | 우울함 수준 | 0~1 |
| surprise | 놀라움 수준 | 0~1 |
| prob\_emotion | 예측 감정 수준 | 0~1 |
| pred\_emotion | 예측 감정 |  |

* 최종 데이터 형태:

**3. 분석 결과(Results)와 시사점**

**3.1 EDA**

* 수치형 변수들 간 상관관계:  
  텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

  자동 생성된 설명

① 음원의 유명세(popularity)는 음원의 특징(audio\_features) 및 가사의 감정과 특별한 상관성을 보이지 않았다.  
② anger – speechiness (0.22): 가사의 분노감 수준이 높을수록 음원의 발성 비중이 증가하는 경향이 약간 보인다.

③ disgust – speechiness (0.24): 가사의 혐오감 수준이 높아질수록 음원의 발성 비중이 증가하는 경향이 약간 보인다.

④ disgust – danceability (0.28): 가사의 혐오감 수준이 높아질수록 음원의 음원의 춤을 추기에 적합한 정도가 증가하는 경향이 약간 보인다.

⑤ added\_year – duration\_ms (-0.30): 해가 지날수록 Top Hit에 수록되는 음원들의 길이가 감소하는 경향이 약간 보인다.

* Top Hit 차트에 진입한 음원의 75% 이상이 금요일에 발매되었다.

텍스트, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

* Top Hit 차트에 진입한 음원의 유명한 정도(popularity)의 분포는 아래와 같다. 최소 65 ~ 최대 95로, 평균값은 78.568이다.

도표, 라인, 그래프, 평면도이(가) 표시된 사진

자동 생성된 설명

* Top Hit 차트에 진입한 음원별 가사에서 가장 많이 등장한 감정은 anger(분노)였다. 부정적 감정(anger, sadness)을 전달하는 음원이 전체 차트 진입 음원의 50% 이상을 차지하고 있다.  
  도표, 스크린샷, 다채로움, 그래프이(가) 표시된 사진

  자동 생성된 설명

아래 상자그림은 음원 가사의 주요 감정별 유명한 정도(popularity)의 분포이다.

도표, 라인, 그래프, 평면도이(가) 표시된 사진

자동 생성된 설명

* 가사에 담긴 감정이 복합적일수록 또는 단순 명확할수록 유명한 정도(popularity)가 달라질까?  
  prob\_emotion의 값이 커질수록 가사의 감정이 단순 명확하며, 값이 작아질수록 가사의 감정이 복합적이라는 의미로 해석될 수 있다.

텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

위의 산점도를 바탕으로 볼 때, 가사가 전달하는 감정이 복합적인 것과 음원의 유명한 정도(popularity)간의 상관성은 보이지 않는다. 대부분의 Top Hit에 수록된 음원들의 가사에는 다양한 감정이 복합적으로 담긴 것으로 보인다. 감정이 단순 명확(prob\_emotion > 0.5)한 음원은 28개에 불과하다.

* 음원의 주요 감정으로 가장 빈번하게 등장한 anger 다음으로, 어떤 감정이 클 때 유명한 정도가 달라질까? 다시 말해 anger가 어떤 감정과 주요 조합을 이룰 때 더 유명할까? 이를 탐색하기 위해 가사의 주요 감정(pred\_emotion)이 anger인 데이터에 한해서, anger 다음으로 감정 score가 큰 감정(second\_emotion)과 해당 score값(second\_prob)을 구한 뒤 유명도(popularity)의 분포를 분석하였다.

스크린샷, 도표, 텍스트, 직사각형이(가) 표시된 사진

자동 생성된 설명

anger와 가장 많은 조합을 이루는 감정은 disgust(혐오감)이었다.

도표, 직사각형, 사각형, 디자인이(가) 표시된 사진

자동 생성된 설명

anger 다음의 주요 감정이 joy일 때, Top hit 중에서도 상위 랭크에 오른 것을 확인할 수 있다.

추가로 sadness의 경우도 살펴보았다.

스크린샷, 텍스트, 도표, 직사각형이(가) 표시된 사진

자동 생성된 설명도표, 직사각형, 사각형, 디자인이(가) 표시된 사진

자동 생성된 설명

sadness와 가장 많은 조합을 이루는 감정은 anger(분노)였다. sadness 다음의 주요 감정이 surprise일 때, Top hit 중에서도 상위 랭크에 오른 것을 확인할 수 있다.

**3.2 Clustering**

* Pycaret의 autoML 라이브러리를 이용한 kmeans 모델링

PyCaret은 파이썬 기반의 오픈소스 머신 러닝 라이브러리로, 머신 러닝 모델의 빠른 실험 및 모델 선택, 하이퍼파라미터 튜닝, 모델 파이프라인 생성 등을 지원한다. PyCaret을 이용해 kmeans 모델을 만들어 500개의 음원 데이터에 대한 군집화를 진행했다.

라인, 텍스트, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 그래프, 도표이(가) 표시된 사진

자동 생성된 설명

텍스트, 스크린샷, 도표, 원이(가) 표시된 사진

자동 생성된 설명**텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명**

3개의 군집이 형성되었으며, 실루엣 계수는 0.5350로, 위 그래프에서 확인할 수 있듯이 적절하게 군집화가 되었다고 볼 수 있다. 각각의 군집의 음원 수는 207, 53, 240개이다.  
 군집별로 변수를 비교해보았으나, 음원의 길이(duration\_ms)를 제외하고 특별한 차이가 두드러진 않았다.

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

직사각형, 도표, 스크린샷, 사각형이(가) 표시된 사진

자동 생성된 설명도표, 스크린샷, 직사각형, 라인이(가) 표시된 사진

자동 생성된 설명

스크린샷, 텍스트, 다채로움, 도표이(가) 표시된 사진

자동 생성된 설명 텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명

\*첨부된 그래프 이외 비교는 BML\_pyacret.ipynb에서 제시되어있다

**3.4 시사점**

본 연구를 통해 스트리밍 음원 사이트의 Top 100 차트에 진입하는 음원들의 내재적 특성을 살펴보았다. 음원의 내재적 특성으로 정형 데이터인 오디오 특징과 비정형 데이터인 가사의 특징을 함께 고려하였다. 일부(28) 음원을 제외한 대다수 음원은 가사에 복합적인 감정을 담고 있었다.

음원의 유명한 정도(popularity)에 대한 분석을 종합해봤을 때, 차트에 진입하는 동력은 음원의 내재적 특성에 있다고 보기 어렵다. 즉 노래만 좋다고 해서 높은 성과를 거둘 수는 없다는 것이다. 음원의 흥행은 음원의 내재적 특성과 외부적 요인(마케팅, 아티스트의 팬덤 등)과의 상호작용에 더 큰 영향을 받는 것으로 예상된다. 따라서 후속 연구로, 음원에서 비롯되는 감정과 리뷰 기반 대중 반응 정서의 일치 정도와 음원의 흥행 대한 연구를 진행할 예정이다.

**4. python code, 참고문헌 및 자료**

* python code:<https://github.com/suachandra/p23_BML.git>  
  BML\_predict\_emotion\_of\_lyrics.ipynb : BERT 모델  
  ANALYSIS\_5years\_HIT100.ipynb: 분석결과1  
  BML\_pycaret.ipynb: 분석결과2
* Rajendran, R. V., Pillai, A. S., & Daneshfar, F. (2022). LyBERT: Multi-class classification of lyrics using Bidirectional Encoder Representations from Transformers (BERT)
* Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *Ismir* (pp. 619-624).ISO 690
* Jochen Hartmann, "Emotion English DistilRoBERTa-base". https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.
* 윤경섭, 오종민.(2022).딥러닝 모델(BERT)과 감정 어휘 사전을 결합한 음원 가사 감정 분석.한국컴퓨터정보학회 학술발표논문집,30(2),471-474.