앙상블 기법을 이용한 한국어 노래 가사의 감정 분류

장동훈, 심규성, 최동희, 박대영
인하대학교 정보통신공학과

ehdgnsdl9192@gmail.com, sgs9712@hanmail.net, dod7000@naver.com, dpark@inha.ac.kr

Emotional Classification of Korean Song Lyrics based on Ensemble Techniques

Donghun Jang, Gyuseong Sim, Donghee Choi, Daeyoung Park
Department of Information and Communication Engineering, Inha University

요 약

인공지능을 기반으로 노래 가사에 담긴 감정을 분류하여 사용자의 음악 취향에 맞춤으로 노래를 추천할 수 있다. 본 논문에서는 타 언어에서의 감정 분류 성능이 좋았던 학습 모델들을 한국어 노래 가사에 적용하여 감정 분류의 성능을 확인한다. 또한 여러 학습 모델의 분류 결과를 합쳐서 분류를 하는 Voting 기반의 앙상블 기법을 사용하여 가장 좋았던 단일 모델의 성능을 뛰어넘는 감정 분류 결과를 얻는다.

Ⅰ. 서론

음악 스트리밍 애플리케이션이 활발하게 사용되면서 음악 추천을 위해 음악의 감정을 분류하는 것이 중요해지고 있다[1]. 빠르게 변화하는 음악 시장에 따라 개별 사용자가 선호하는 장르의 음악을 찾는 일이 많아지면서 각종 음원 사이트에서는 사용자 맞춤 음악 장르 추천이 활성화되고 있다. 인공지능을 이용한 정확한 감정 분류가 사용자 맞춤형으로 음악 추천을 하는 데 도움을 줄 것이다.

기존 문헌에서는 영어 노래 가사 데이터 셋에 감정의 범주를 차원형 Russell 모델의 네 가지 감정 범주로 분류했다[2]. 또한 데이터 전처리 중 하나인 글로브(GloVe) 방법론을 적용한 Bi-LSTM(Bidirectional Long-Short Term Memory) 딥러닝 방법을 사용하여 곡의 감정을 분류했다[3]. 그리고 Turkey 언어에서 감정 분류 성능이 좋았던 CLDNN(Convolutional Long Short Term Memory Deep Neural Network) 모델을 한국어 노래 가사에 적용함으로써 감정 분류 성능을 확인 및 개선하였다[4]. 또한 ML(Machine Learning) 및 DNN(Deep Neural Network) 모델들 외에 Bengali 언어에서 성능이 좋았던 Transformer 모델을 추가함으로써 감정 분류를 하였다[5].

본 논문에서는 타 언어에서 좋았던 모델들을 한국어 노래 가사 데이터 셋에 적용하여 성능을 확인한다. 음악 서비스 플랫폼인 벅스에서 웹 크롤링을 이용하여 한국어 노래 가사 데이터를 수집한다. 그리고 타 언어에서의 감정 분류 성능이 좋았던 ML, DNN, Transformer 모델들을 구현하여 한국어에서의 감정 분류 성능을 확인하고 비교한다. 또한 단일 모델의 분류 결과를 모아서 분류를 수행하는 Voting 앙상블 기법을 사용하여 감정 분류 성능을 높이는 방법을 제안한다. 각 모델의 데이터 전처리 방법 및 결과 분석을 설명하며 감정 분류 성능을 높인 기술에 관해 서술한다.

Ⅱ. 노래의 감정 분류 방법

2-1. 데이터 수집 및 정제

음원 서비스 플랫폼 벅스에서 웹 크롤링 (Web Crawling)을 이용하여 클래스 비율을 균일하게 총 12,800 개의 한국어 노래 가사 데이터 셋을 수집했다. 벅스에서는 '감정/기분'이라는 범주로 구분돼 있기 때문에 차원형 Russell 모델에 의거하여 총 4 가지 Happy, Sad, Relaxed, Angry 감정으로 분류하여 라벨로 정의했다. 그리고 훈련 데이터 세트 수는 7,168, 검증 데이터 세트 수는 3,072, 테스트 데이터 세트 수는 2,560 으로 나누어서 실험을 진행했다.

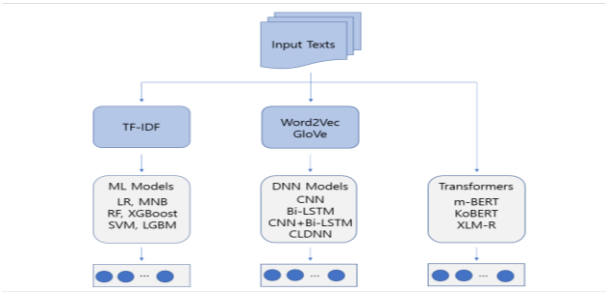

그림 1. 감정 분류의 과정

그림 1 은 각 감정 분류 모델에서 데이터 전처리 기법과 감정 분류 모델을 전체적으로 보여준다. 데이터 정제 과정으로 한글 외에 나머지 문자들은 모두 제거하였다. 또한, 한국어 조사 등의 문제로 TF-IDF 및 Word2Vec 에 적용하기 이전에 토큰화 및 POS 태깅, 불용어 처리를 한다. 하지만, 한국어에서 불용어 처리를 하지 않은 경우가 처리한 경우보다 더 성능이 좋았으며 불용어 처리 여부보다 형태소로 나누는 토큰 단위가 더 중요하다[6]. 본 연구에서는 정확하고 빠르게 형태소를 나눠주는 Twitter 계열의 Okt 형태소 분석기를 사용했다. 그리고 불용어 처리를 하지 않은 경우가 불용어 처리한 경우보다 성능이 좋았기 때문에 불용어 처리를 하지 않았다.

2-2. 각 모델의 구현 방법

데이터 전처리로 ML 모델에서는 TF-IDF 기법, DNN 모델에서는 Word2Vec과 GloVe 기법을 사용하였다. Transformer 모델에서 사용되는 텍스트 토크나이저인 SentencePiece는 영어와 다르게 한국어의 띄어쓰기에 따른 토큰화가 어려운 점을 해결하여 한국어의 변칙적인 언어 특성을 반영하기 위해 데이터 기반의 토큰화

기법을 적용한다. 마지막으로 앙상블 모델에서는 앙상블 기법 중 하나인 Voting 방법은 투표를 통해 값을 결정하는 방법으로 Hard Voting과 Soft Voting 방법이 있다[7]. Hard Voting은 여러 분류기가 예측한 값을 최종 값으로 선택하는 방법이고, Soft Voting은 각 레이블끼리 확률 예측을 하고 평균을 내어 최종 값을 결정하는 방법이다. 단일 ML 모델 SVM(Support Vector Machine), LR(Linear Regression), RF(Random Forest), MNB(Multinomial Naive Bayes), XGBoost(Extreme Gradient Boosting), LGBM(Light Gradient Boosting Model) 중 2~6개의 모델을 선택하여 Soft Voting과 Hard Voting의 앙상블 기법을 사용해 모델을 구현한 후 두 Voting 방법 간에 성능 차이를 비교하였다.

Ⅲ. 실험 결과

한국어 가사 감정을 분류하기 위해 다양한 ML, DNN, Transformer 모델들의 성능 분석을 제공한다. 모델의 성능은 가중된 F1 Score 로 결정되지만 추가로 Precision, Recall, Accuracy, Area under the Curve 도 고려되었다.

표 1. 모든 모델의 성능을 비교한 최종 결과

Method	Classifier	Accuracy	Precision	Recall	F1	AUC
ML models	LR	74.10	74.13	74.07	74.07	82.72
	SVM	74.14	74.21	74.10	74.10	72.74
	RF	69.57	71.31	69.72	69.65	79.81
	MNB	73.82	74.19	73.74	73.83	82.51
	XGBoost	71.56	72.45	71.66	71.64	81.11
	LGBM	71.83	72.42	71.88	71.89	81.26
	Ensemble	**75.16**	**75.40**	**75.10**	**75.16**	**83.41**
DNN models	Bi-LSTM (Word2Vec)	63.07	63.28	63.12	63.17	75.41
	Bi-LSTM (GloVe)	61.48	62.24	61.50	61.48	74.33
	CNN+Bi-LSTM (Word2Vec)	67.30	67.22	66.31	67.20	78.20
	CNN+Bi-LSTM (GloVe)	57.70	57.60	57.63	57.85	71.77
Trans-formers	XLM-R	72.30	72.81	72.41	72.41	81.59
	KoBERT	74.22	74.74	74.26	74.35	82.83

표 1 은 한국어 노래 가사 감정 분류를 수행하는 모든 모델의 성능을 비교한 결과이다. ML 모델 중 SVM 모델의 성능이 좋았으며, DNN 모델 중 Word2Vec 으로 전처리 CNN+Bi-LSTM 모델이 좋았으며, Transformer 모델 중 KoBERT 모델의 성능이 좋았다. ML, DNN, Transformer 모든 모델을 비교한 단일 모델에서는 Transformer 계열의 KoBERT 모델의 성능이 가장 우수했다.

또한 성능 향상을 위해 앙상블 기법의 하나인 Voting 앙상블 기법을 사용했다. 결과적으로, 앙상블 모델 중 성능 상위 5 개 단일 ML 모델을 가지고 Soft Voting 앙상블 기법을 사용한 모델이 가장 우수했다. 제안된 앙상블 모델이 단일 모델에서 가장 좋았던 Transformer 계열의 KoBERT 모델 대비, Accuracy 는 1.17%, Precision 는 1.26%, Recall 는 1.21%, F1 Score 는 1.22%, AUC 는 0.80%의 성능 향상을 보였다. 제안된 앙상블 모델을 사용함으로써 단일 모델 중 감정 분류 성능이 가장 높게 나왔던 KoBERT 모델의 성능을 능가했다.

Ⅳ. 결론

기존 연구에서는 타 언어에서의 감정 분류를 위하여 ML, DNN, Transformer 모델들을 모두 구현하여 비교한 경우가 적었고 한국어 노래 가사 데이터를 사용하여 모든 모델을 구현하여 감정 분류 성능을 비교한 경우가 적었다. 본 논문에서는 타 언어에서 감정 분류 성능이 좋았던 모델들을 사용하여 한국어에서의 감정 분류 성능을 확인하고 비교했다. 여러 모델의 성능을 합친 앙상블 기법으로 단일 모델의 성능보다 우수한 감정 분류 성능을 달성하였다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2022-00155915, 인공지능융합혁신인재양성(인하대학교)). 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1H1A2092541).

참 고 문 헌

[1] R. Rajendran, A. Pillai, and F. Daneshfar, “Multi-class classification of lyrics using Bidirectional Encoder Representations from Transformers (BERT),” Hindustan Institute of Technology and Science, pp. 2-14, 2022.

[2] E. Çano and M. Morisio, “MoodyLyrics: A sentiment annotated lyrics dataset,” International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, Hong Kong, pp. 118-124, 2017.

[3] S. Hizlisoy, S. Yildirim, and Z. Tufekci, “Music emotion recognition using convolutional long short term memory deep neural networks,” Engineering Science and Technology, an International Journal, pp. 760-767, 2021.

[4] J. Abdillah, I. Asror, Y. Firdaus, and A. Wibowo, “Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting,” System Engineering and Information Technology, vol. 4, no. 4, pp. 723-729, 2020.

[5] A. Das, O. Sharif, M. Hoque, and I. Sarker, “Emotion classification in a resource constrained language using transformer-based approach,” Association for Computational Linguistics, pp. 150-158, 2021.

[6] Y. S. Jeong, T. S. Heo, and Y. S. Kim, “한국어 감성 분석에서 토큰 단위와 불용어 처리 기준에 따른 성능 비교,” 한국정보과학회 학술발표논문집, pp. 1394-1396, 2019.

[7] M. Salur and I. Aydin, “A soft voting ensemble learning-based approach for multimodal sentiment analysis,” Neural Comput & Appl., pp. 18391-18406, 2022.