Music Recommendation Using PLM sentiment classification

14팀 계원재 김두현 신인섭 장건호

1. 서론

현재 음악 추천 서비스에서의 “가사”의 위치는 어디인가를 시작으로 이번 프로젝트를 진행하였다.
Spotify의 Web API를 통해서 추천 알고리즘을 살펴보면, 사용자가 노래를 선택시, 해당 노래를 들은 사용자들 중 재생회수의 thresholding을 하여서 1차적으로 사용자 후보군을 정한다. 이것을 바탕으로 각 사용자가 들은 노래들을 순서데로 나열하고, K-NN을 활용해서 최적의 음악을 추천한다.
이것과 더불어 NLP를 활용하여서 해당 노래의 기사나 가수의 article을 바탕으로 description을 작성하고 vector화 한다고 나와있다.

해당 과정에서 노래 자체의 lyrics에 대한 부제를 확인하였고, 이를 활용하면 다음과 같은 이득이 있을 것이라고 판단하였다. 먼저, 사용자 로그 기반 추천은 아무래도 시간적인 제약에 갇힐 수 밖에 없는 알고리즘이다. 이를 가사를 활용하면, 흔히 말하는 음악 digging이 가능하다고 판단하였다. 특히 시공간적인 거리가 있는 노래간의 감정선이 비슷한 곡을 추천할 수 있다면, 새로운 추천이 가능하다고 생각하였다.

2. 영어 가사 데이터셋(Baseline)

1. go_emotions 데이터셋 소개

go_emotions 데이터셋은 Reddit의 483개 커뮤니티에서 추출된 댓글로 구성되며, 각 댓글은 하나 이상의 감정 태그와 연결되어 있습니다. 이 데이터셋은 훈련, 검증, 테스트로 나뉘며 각각 43,410, 5,426, 5,427개의 데이터를 포함하고 있습니다. 이를 활용하여 다양한 감정을 인식할 수 있는 모델을 구축할 수 있습니다.

2. QLoRA의 효과적 활용

QLoRA(Quantized Low-Rank Adaptation)는 대규모 언어 모델의 파인튜닝을 효율적으로 수행할 수 있도록 돕는 기술입니다. 이 기법은 모델의 가중치를 4비트로 양자화하여 메모리 사용량을 감소시키면서도 성능 저하를 최소화합니다. 특히, 자원이 제한된 환경에서도 고효율의 파인튜닝이 가능하여 대규모 모델을 경제적으로 활용할 수 있게 합니다.

3. 모델 학습과 성능 비교

RoBERTa 모델을 QLoRA와 결합하여 4,000개의 샘플 데이터로 학습한 결과, 메모리 관리와 학습 시간이 효율적으로 개선되었습니다. 이러한 설정에서 QLoRA를 사용한 경우(RoBERTa with QLoRA)가 사용하지 않은 경우(RoBERTa without QLoRA)와 비교하여 유사한 예측 성능을 보였습니다.

4. 노래 추천 시스템

roberta-base-go_emotions 모델을 사용하여 노래 가사의 감정을 분석하고, 이를 통해 사용자의 감정 상태에 맞는 노래를 추천하는 시스템을 구축했습니다. 이 시스템은 노래 가사를 특수 기호에서 정제하고, 토큰화 및 text-classification을 통해 각 곡을 대표하는 상위 3개의 감정 태그를 추출합니다. 그 후, 코사인 유사도(cosine similarity)를 사용하여 감정 태그가 유사한 다른 곡들을 추천합니다.

3. 한국어 감정 분석 모델 분석

1. 한국어 가사 dataset의 부재

한국어의 경우 감성 태깅된 대규모 가사 데이터셋의 부재가 가장 큰 걸림돌이였다. 따라서 본 프로젝트에서는 가사를 기반으로 학습을 하기보다는 감성 분석의 foundation model의 성능을 향상 시키는 것에 집중을 하고, 해당 모델을 한국어 가사에 naive하게 적용하여 모델의 성능과 한계를 분석한다.

2. 한국어 기반 pre-trained Model

한국어에 대해 감성분석 모델을 구성하기 위해서 pre-trained 모델로 klue/bert-base를 채택하였다. 해당 모델은 아래과 같은 데이터를 학습한 상태이다.

- MODU: Modu Corpus is a collection of Korean corpora distributed by National Institute of Korean Languages. It includes both **formal articles (news and books) and colloquial text (dialogues).**
- CC-100-Kor: CC-100 is the large-scale multilingual web crawled corpora by using CC-Net (Wenzek et al., 2020). This is used for training XLM-R (Conneau et al., 2020). We use the Korean portion from this corpora.
- NAMUWIKI: **NAMUWIKI** is a Korean web-based encyclopedia, similar to Wikipedia, but known to be less formal. Specifically, we download the dump created on March 2nd, 2020.
- NEWSCRAWL: NEWSCRAWL consists of 12,800,000 news articles published from 2011 to 2020, collected from a **news aggregation platform**.
- PETITION: Petition is a collection of public petitions posted to the Blue House asking for administrative actions on social issues. We use the **articles in the Blue House National Petition** published from August 2017 to March 2019.

위의 총 62GB의 데이터를 학습시켰으며, KLUE는 한국어 NLU에 대한 bechmark를 제공하고 있다.

KLUE Leaderboard

Unlike other benchmarks, klue benchmarks do not provide total scores and leaderboards for the entire task. On the leaderboard, you can check each score for one model and sort by each evaluation metric.

All | Small Size | Base Size | Large Size

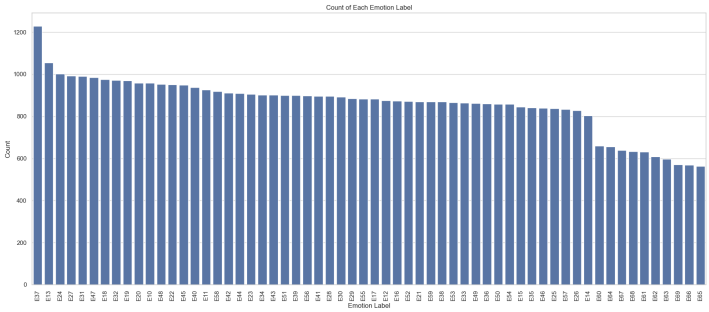
#	Team	Model	Description	YNAT	KLUE-STS		KLUE-NLI	KLUE-NER		KLUE-RE		KLUE-DP		KLUE-MRC		WOS	
				F1	R^P	F1	ACC	$F1^E$	$F1^C$	$F1^{MK}$	AUC	UAS	LAS	EM	ROUGE	JGA	$F1^S$
1	KLUE-team	KLUE-BERT-base	More	85.73	90.85	82.84	81.63	83.97	91.39	66.44	66.17	89.96	88.05	62.32	68.51	46.64	91.61
2	KLUE-team	KLUE-RoBERTa-large	More	85.69	93.35	86.63	89.17	85	91.86	71.13	72.98	93.48	88.36	75.58	80.59	50.22	92.23
3	KLUE-team	KLUE-RoBERTa-base	More	85.07	92.5	85.4	84.83	84.6	91.44	67.65	68.55	93.04	88.32	68.67	73.98	47.49	91.64
4	KLUE-team	KLUE-RoBERTa-small	More	84.98	91.54	85.16	79.33	83.65	91.14	60.89	58.96	90.04	88.14	57.32	62.7	46.62	91.44
5	KLUE-tester		More	79.63	88.51	81.22	67.03	81.07	89.39	44.86	31.99	89.58	88.03	40.74	45.86	2.44	48.04

벤치마크는 주제 분류, 의미적 텍스트 유사성, 자연어 추론, 개체명 인식, 관계 추출, 의존성 파싱, 기계 독해 이해, 대화 상태 추적으로 구성된다.

3. 감성 분석 데이터셋

Klue/bert-base PLM을 이용해서 감정 분석을 위한 모델을 만들기 위해서 ai-hub에서 제공하는 “감성 대화 말뭉치”을 이용한다. 이 데이터셋의 원시데이터는 크라우드 소싱 방식을 통해 수집하며, 피험자 개개인이 주어진 상황에 맞는 데 이터를 직접 입력한다. 개인별 주어진 상황에 대한 감정 상태의 표현을 기재하므로, 수집 단 계에서 특별한 공통 관리가 필요하지는 않다. 해당 데이터셋의 목적은 AI 기반 감성 챗봇용 세대별 감성대화 텍스트 데이터의 구축이며, 감성 라벨을 총 60개로 구분되어 있다.

4. Dataset EDA



전반적으로 고른 분포를 보이지만, 감정 태깅 자체가 부정적인 감정들에 편향 되어 있다. 총 60가지 중에 긍정에 해당하는 감정을 1/6밖에 되지 않는다. 그럼에도 워낙 데이터셋의 양이 크기 때문에(trianing 14만개의 corpus) 특별한 분리 없이 모든 데이터를 사용하여 학습을 진행하였다.

5. hun3359/klue-bert-base-sentiment

본 감성분석 fine-tunning에 앞서서 huggingface에 동일한 데이터셋을 바탕으로 fine-tunning 되어 있는 모델을 확인하였다.

해당 모델에 가사를 바로 적용한 예시들은 아래와 같다.

쿨 - 애상

Top 1 Prediction: 악의적인 (Probability: 0.0743)
Top 2 Prediction: 안달하는 (Probability: 0.0526)
Top 3 Prediction: 분노 (Probability: 0.0525)
Top 4 Prediction: 짜증내는 (Probability: 0.0494)
Top 5 Prediction: 노여워하는 (Probability: 0.0427)
Top 6 Prediction: 희생된 (Probability: 0.0400)
Top 7 Prediction: 염세적인 (Probability: 0.0398)
Top 8 Prediction: 성가신 (Probability: 0.0327)
Top 9 Prediction: 구역질 나는 (Probability: 0.0315)
Top 10 Prediction: 방어적인 (Probability: 0.0282)

칵테일 사랑

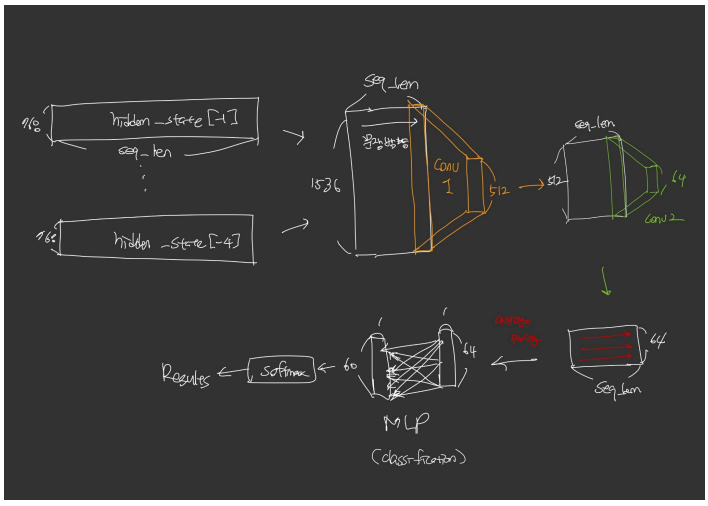
Top 1 Prediction: 악의적인 (Probability: 0.0599)
Top 2 Prediction: 두려운 (Probability: 0.0486)
Top 3 Prediction: 안달하는 (Probability: 0.0483)
Top 4 Prediction: 성가신 (Probability: 0.0401)
Top 5 Prediction: 분노 (Probability: 0.0325)
Top 6 Prediction: 노여워하는 (Probability: 0.0289)
Top 7 Prediction: 취약한 (Probability: 0.0268)
Top 8 Prediction: 방어적인 (Probability: 0.0267)
Top 9 Prediction: 염세적인 (Probability: 0.0262)
Top 10 Prediction: 희생된 (Probability: 0.0262)

성시경 - 거리에서

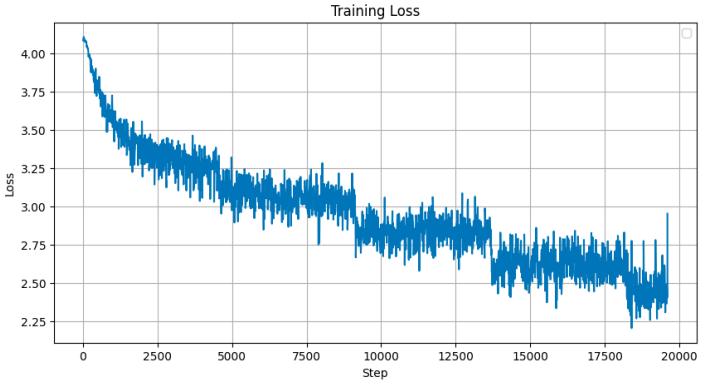
Top 1 Prediction: 두려운 (Probability: 0.0552)
Top 2 Prediction: 안달하는 (Probability: 0.0447)
Top 3 Prediction: 악의적인 (Probability: 0.0406)
Top 4 Prediction: 성가신 (Probability: 0.0386)
Top 5 Prediction: 분노 (Probability: 0.0368)
Top 6 Prediction: 염세적인 (Probability: 0.0325)
Top 7 Prediction: 방어적인 (Probability: 0.0317)
Top 8 Prediction: 노여워하는 (Probability: 0.0316)
Top 9 Prediction: 혼란스러운 (Probability: 0.0307)
Top 10 Prediction: 취약한 (Probability: 0.0263)

전반적으로 정확하지 못한 성능을 보이는 것을 볼 수 있다.
이는 **긴 text에 대해서는 학습이 잘 안되었다고 느껴졌다**.

6. Architecture 1.



conv1d를 두번 사용하는 구조이다.
padding=1, kernel size=2로 두번에 걸쳐서 차원을 64로 낮춘다.
그후 average pooling 후, MLP로 60개에 대한 classifier를 만든다.



Training loss

<가사 대입 예시>
쿨 - 애상
Top 1 Prediction: 질투하는 (Probability: 0.0813)
Top 2 Prediction: 배신당한 (Probability: 0.0412)
Top 3 Prediction: 악의적인 (Probability: 0.0366)
Top 4 Prediction: 염세적인 (Probability: 0.0363)
Top 5 Prediction: 외로운 (Probability: 0.0342)
Top 6 Prediction: 상처 (Probability: 0.0338)
Top 7 Prediction: 충격 받은 (Probability: 0.0308)
Top 8 Prediction: 고립된 (Probability: 0.0307)
Top 9 Prediction: 버려진 (Probability: 0.0285)
Top 10 Prediction: 회의적인 (Probability: 0.0279)

칵테일 사랑
Top 1 Prediction: 우울한 (Probability: 0.1286)
Top 2 Prediction: 염세적인 (Probability: 0.0847)
Top 3 Prediction: 슬픔 (Probability: 0.0754)
Top 4 Prediction: 눈물이 나는 (Probability: 0.0573)
Top 5 Prediction: 비통한 (Probability: 0.0467)
Top 6 Prediction: 회의적인 (Probability: 0.0325)
Top 7 Prediction: 후회되는 (Probability: 0.0315)
Top 8 Prediction: 고립된 (Probability: 0.0285)
Top 9 Prediction: 낙담한 (Probability: 0.0260)
Top 10 Prediction: 고립된(당황한) (Probability: 0.0260)

성시경 - 거리에서
Top 1 Prediction: 눈물이 나는 (Probability: 0.0960)
Top 2 Prediction: 슬픔 (Probability: 0.0801)
Top 3 Prediction: 우울한 (Probability: 0.0723)
Top 4 Prediction: 마비된 (Probability: 0.0601)
Top 5 Prediction: 비통한 (Probability: 0.0360)
Top 6 Prediction: 염세적인 (Probability: 0.0343)
Top 7 Prediction: 낙담한 (Probability: 0.0338)
Top 8 Prediction: 좌절한 (Probability: 0.0320)
Top 9 Prediction: 후회되는 (Probability: 0.0288)
Top 10 Prediction: 실망한 (Probability: 0.0262)

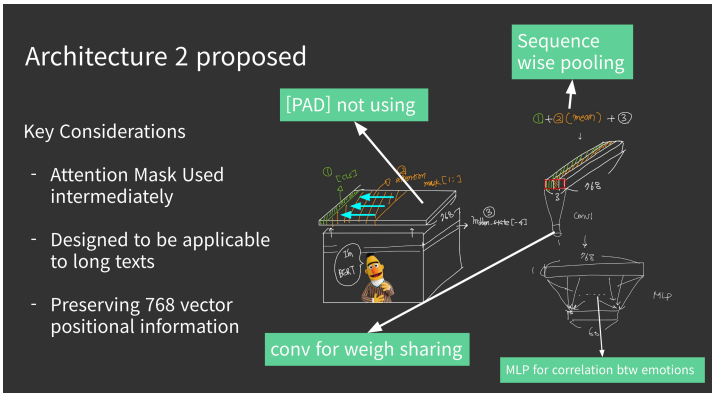
해당 예시들에는 유의미한 결과를 가져오는 것 같지만, 조금 긍정적인 가사인 싸이 - 예술이야 의 결과는
Top 1 Prediction: 신이 난 (Probability: 0.0365)
Top 2 Prediction: 질투하는 (Probability: 0.0364)
Top 3 Prediction: 구역질 나는 (Probability: 0.0359)
Top 4 Prediction: 흥분 (Probability: 0.0347)
Top 5 Prediction: 혐오스러운 (Probability: 0.0300)
Top 6 Prediction: 부끄러운 (Probability: 0.0293)
Top 7 Prediction: 노여워하는 (Probability: 0.0285)
Top 8 Prediction: 눈물이 나는 (Probability: 0.0284)
Top 9 Prediction: 우울한 (Probability: 0.0281)
Top 10 Prediction: 스트레스 받는 (Probability: 0.0263)

싸이 예술이야에 대해서 감성 분석을 돌리면, “신이 난” top1 결과는 잘 나오지만, top2-5에서 상당히 부정적인 단어들이 많이 등장하는 것을 볼 수 있다.
이는 top1을 신이 난이라고 예측하면서도 바로 밑 예측이 굉장히 거리감이 있단느 사실이 특징적이다. 이는 “모델이 감정간의 유사성을 전혀 학습하지 못했다”라는 결론을 내렸다. 이를 아래와 같은 이유로 분석해 보았다.

6-1. Architecture 1 의 취약점

Conv1d를 활용한 모델에서는 output으로 나오는 768개의 feature를 64개로 줄였다. 여기서 1차적으로 bert가 의미있게 만들어낸 output을 뭉갰다고 생각하였다. 그리고 그 다음에 seq_len 방향으로의 학습이 아니라, average pooling을 했다는 점은 문장 간의 연과성이나 감정 간의 연결을 blurring한 것과 같다고 판단하였다.
따라서 해당 문제점을 극복하기 위해서 아래와 같은 새로운 architecture을 제안하였다.

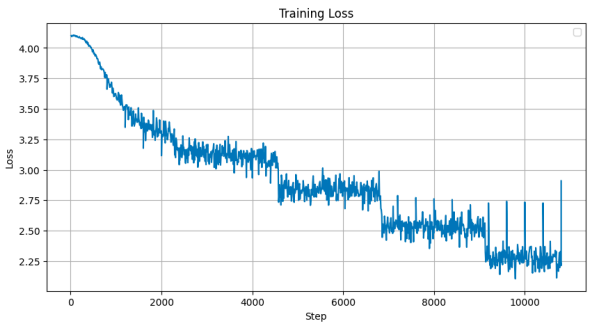
7. Architecture 2.



여기서 특별히 앞서 나타났던, 문제점을 고려하기 위해서 추가적인 구조를 마련하였다.
먼저, cls token의 last output과 [-4] output을 활용하는 것은 동일하다.
… (1)
** 이때, 우리가 학습하는 데이터는 대부분 문장이 1개인 데이터 셋이다.**
따라서 나는 여기서 가사라는 긴 문장에 대해서도 모델이 대응이 가능하도록 attention mask를 활용한 학습을 진행하였다. (1)에 더불어서 [PAD] 부분을 output을 제외한, attention mask가 걸린부분만 classification layer에서 활용하도록 하였으면, 해당 (seq_len-1, 768) 차원을 (1, 768)로 만들기 위해서 mean pooling을 진행하였다.

```
attention_mask_expanded = attention_mask.unsqueeze(-1).expand_as(hidden_states[-1])
masked_last_layer = hidden_states[-1][:, 1:, :] * attention_mask_expanded[:, 1:, :]
mean_pooled_vector = masked_last_layer.mean(dim=1)
```

그리고 (1)과 concat하여 (3,768) tensor를 만들었다. 해당 tensor를 conv1d를 통해서 (1,768)로 만들었으면, 해당 768vector를 60개로 mapping하는 linear layer를 사용하여 classifier를 만들었다.



아래는 해당 모델의 가사 예시이다.
쿨 - 애상
Top 1 Prediction: 외로운 (Probability: 0.2201)
Top 2 Prediction: 안달하는 (Probability: 0.0580)
Top 3 Prediction: 두려운 (Probability: 0.0511)
Top 4 Prediction: 질투하는 (Probability: 0.0496)
Top 5 Prediction: 눈물이 나는 (Probability: 0.0482)
Top 6 Prediction: 염세적인 (Probability: 0.0425)
Top 7 Prediction: 배신당한 (Probability: 0.0290)
Top 8 Prediction: 상처 (Probability: 0.0237)
Top 9 Prediction: 버려진 (Probability: 0.0227)
Top 10 Prediction: 고립된(당황한) (Probability: 0.0221)

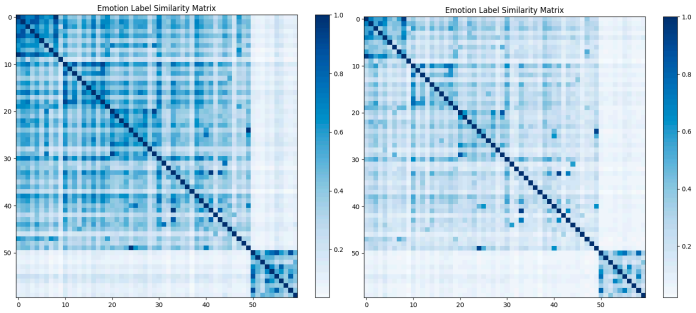
칵테일 사랑
Top 1 Prediction: 우울한 (Probability: 0.2871)
Top 2 Prediction: 염세적인 (Probability: 0.0808)
Top 3 Prediction: 가난한, 불우한 (Probability: 0.0527)
Top 4 Prediction: 눈물이 나는 (Probability: 0.0455)
Top 5 Prediction: 슬픔 (Probability: 0.0430)
Top 6 Prediction: 비통한 (Probability: 0.0377)
Top 7 Prediction: 낙담한 (Probability: 0.0249)

Top 8 Prediction: 환멸을 느끼는 (Probability: 0.0246)
Top 9 Prediction: 마비된 (Probability: 0.0215)
Top 10 Prediction: 회의적인 (Probability: 0.0190)

성시경 - 거리에서
Top 1 Prediction: 눈물이 나는 (Probability: 0.2315)
Top 2 Prediction: 마비된 (Probability: 0.1358)
Top 3 Prediction: 염세적인 (Probability: 0.0941)
Top 4 Prediction: 슬픔 (Probability: 0.0832)
Top 5 Prediction: 우울한 (Probability: 0.0518)
Top 6 Prediction: 비통한 (Probability: 0.0348)
Top 7 Prediction: 억울한 (Probability: 0.0267)
Top 8 Prediction: 환멸을 느끼는 (Probability: 0.0261)
Top 9 Prediction: 고립된 (Probability: 0.0242)
Top 10 Prediction: 후회되는 (Probability: 0.0205)

싸이 - 예술이야
Top 1 Prediction: 흥분 (Probability: 0.5096)
Top 2 Prediction: 신이 난 (Probability: 0.1202)
Top 3 Prediction: 기쁨 (Probability: 0.0586)
Top 4 Prediction: 자신하는 (Probability: 0.0410)
Top 5 Prediction: 마비된 (Probability: 0.0295)
Top 6 Prediction: 눈물이 나는 (Probability: 0.0292)
Top 7 Prediction: 만족스러운 (Probability: 0.0152)
Top 8 Prediction: 느긋 (Probability: 0.0134)
Top 9 Prediction: 안달하는 (Probability: 0.0125)
Top 10 Prediction: 편안한 (Probability: 0.0091)

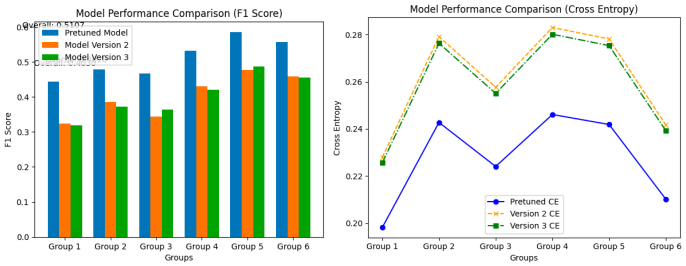
해당 예시들을 보면 확실히 감정간 유사성이 높아진 모습을 볼 수 있다.
조금 더 시각적으로 감정간 유사도를 비교해보면,



확실히 유사도간 긍정과 부정의 correlation이 증가했음을 알 수 있다.

8. Model간 평가지표

본래 classification이라고 하면, f1 score를 떠오르기 마련이다. 하지만, 이는 label가의 dependent가 없는 상황에서는 정확하지만, 지금과 같이 레이블간 가장 correlation이 있는 경우에는 다소 부정확할 것이라는 생각이 들었다. Architecture2와 hugging face fine-tuned model의 f1 score와 cross entropy를 비교해보면 아래와 같다.

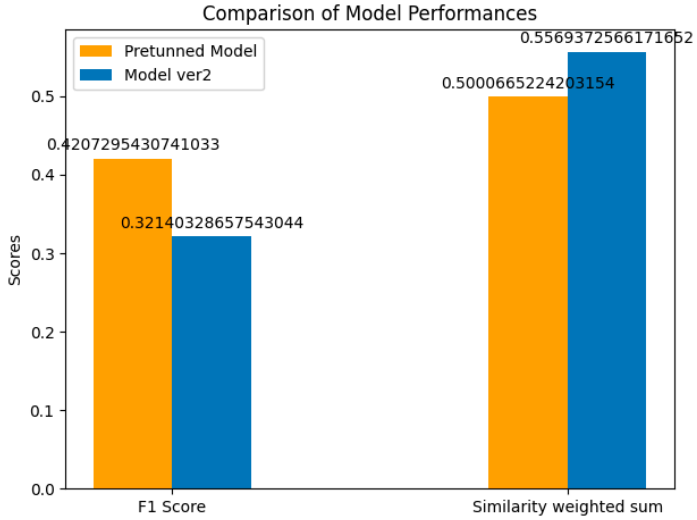


모든 부분에서 structural tunned된 모델이 부족함을 볼 수 있다.

하지만 이것은 perceptual하게 가사의 감성분석의 예시에서 보았던 것과 다르다. 따라서 본 프로젝트에서는 평가지표를 개선해야될 필요성을 느꼈고, 아래와 같은 감정 유사도 기반 평가방법을 소개한다.

$$L = \sum_{i=1}^{N} \left(\sum_{j=1}^{C} S_{y_i,j} \cdot P_{i,j} \right)$$

이 지표는 단순이 True/False만을 따지지 않고, 해당 정답 레이블을 기준으로 60개의 label의 유사도 matirx **S**를 활용하여서, classifier의 ouput matrix **P**의 top1이 아닌 60개의 모든 label에 대해서 weighed sum을 취하는 방식을 활용한다. 이를 통해 평가한 모델의 점수는 아래와 같다.



이 평가지표에서는 Model ver2(Architecture 2)에서 좋은 점수가 나타남을 볼 수 있다.

다음과 같은 평가지표가 시사하는 바는 sentiment analysis와 같은 task의 경우에는 기본적으로 classification의 성향의 task라기 보다는 감정간의 연관서을 얼마나 잘 학습하였냐는 것도 중요한 지표임을 나타낸다. 특히 긴 text에서는 해당 감정의 빈도나 중요도에 따라서 output을 내보내는 것이 아니라 각 감정간의 유사도를 바탕으로 새로운 감정으로의 추론이 가능하도록 만드는 것도 모델이 갖춰야한 능력이라고 판단하다. 따라서 감성분석 task는 평가지표에 있어서도 논의가 되어야 할 것이라는 생각이 든다.

---

4. 발표 후 추가 연구

1) 존재하지 않는 한국어 가사 데이터를 조금이라도 crowling하여서 inference만을 위한 데이터를 확보한다.
2) 해당 데이터를 기반으로 각 노래의 감성 분석을 Architecture2를 이용하여서 적용한다.
3) 해당 곡들의 output probability의 cosine simliarity를 계산한다.
4) 추천 시스템으로서의 역할을 할 수 있을지 여부를 판단한다.

1. 한국어 가사 데이터 crowling

본 프로젝트에서는 chromedriver를 활용해서, songid의 css에 접근한다. 해당 css tag에서 songid를 수집하고 배열화한다.
그 후 songid를 활용하여 해당 노래 가사를 .section_lyric .lyric 태그를 이용하여 수집하고 tuple형태로 저장한다.
일차적으로 “멜론 연도별 1위~10위 곡<1991년~2014년>” 플레이리스트에 대해서 적용하였으며, 237곡을 확보하였다.

이를 선택한 이유는 해당 playlist가 유명한 노래를 중심으로 기본 추천 시스템의 문제점이였던 시간적 거리감을 포함하고 있었기 때문이다.
본 과정은 12분이 소요되었다.

2. 한국어 가사 기반 Architecture 2 적용 inference

썸 (Feat. 릴보이 Of 릭스)

Rank	Prediction	Probability
1	초조한	0.1474
2	안달하는	0.0852
3	걱정스러운	0.0389
4	억울한	0.0383
5	혼란스러운	0.0367
6	불안	0.0359
7	상처	0.0334
8	질투하는	0.0333
9	짜증내는	0.0305
10	방어적인	0.0286

야생화

Rank	Prediction	Probability
1	눈물이 나는	0.3509
2	슬픔	0.1005
3	미비된	0.0835
4	염세적인	0.0430
5	비통한	0.0384
6	우울한	0.0356
7	환멸을 느끼는	0.0282
8	후회되는	0.0185
9	억울한	0.0163
10	낙담한	0.0156

눈, 코, 입

Rank	Prediction	Probability
1	눈물이 나는	0.3250
2	슬픔	0.0835
3	미비된	0.0814
4	비통한	0.0587
5	우울한	0.0336
6	후회되는	0.0314
7	억울한	0.0284
8	상처	0.0256
9	염세적인	0.0233
10	환멸을 느끼는	0.0231

금요일에 만나요 (Feat. 장이정 Of HISTORY)

Rank	Prediction	Probability
1	초조한	0.0917
2	안달하는	0.0839
3	질투하는	0.0591
4	염세적인	0.0403
5	걱정스러운	0.0272
6	외로운	0.0271
7	고립된(당황한)	0.0249
8	느긋	0.0243
9	남의 시선을 의식하는	0.0219
10	우울한	0.0215

한여름밤의 꿀

Rank	Prediction	Probability
1	기쁨	0.3943
2	만도	0.1334
3	신이 난	0.1228
4	흥분	0.1049
5	만족스러운	0.0817
6	느긋	0.0405
7	감사하는	0.0362
8	편안한	0.0297
9	자신하는	0.0194
10	눈물이 나는	0.0120

봄 사람 벚꽃 말고

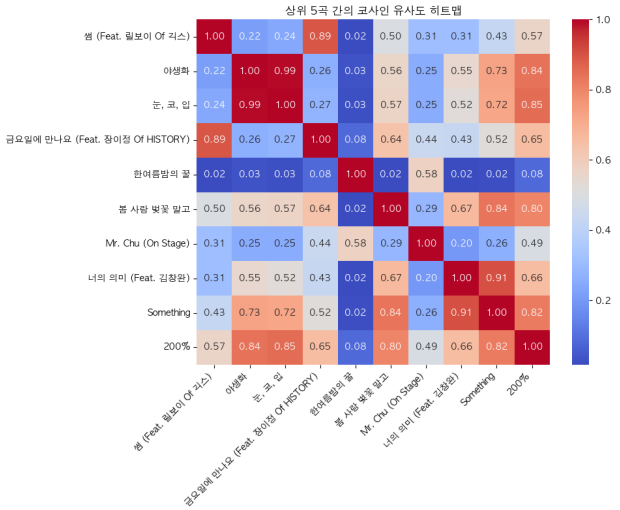
Rank	Prediction	Probability
1	우울한	0.1035
2	눈물이 나는	0.0552
3	미비된	0.0546
4	염세적인	0.0422
5	고립된(당황한)	0.0398
6	슬픔	0.0385
7	외로운	0.0354
8	안달하는	0.0333
9	고립된	0.0321
10	툴툴대는	0.0288

해당 inference는 html 파일로 함께 제공하였습니다.
AI-hub 감성 대화 말뭉치가 부정적인 데이터에 편향되었음에도 불구하고, 썸이나 설램을 생각보다 잘 분석하고 있음을 볼 수 있다. 특히 "한 여름밤의 꿀"에 대한 분석은 놀랍다. 60개 중에 10개 밖에 없는 긍정적인 감정태그가 다수 포함된 것은 확실히 모델이 감정 자체의 유사성을 파악하고 있음을 나타낸다.

이 데이터의 경우, 시대를 넓게 포함하고 있지만 노래의 장르가 대부분 이별에 관한 것이 많음이 확인된다. 시대별 top 10의 노래를 모았기 때문에 이는 대중들이 해당 주제를 좋아하는 것으로 해석되며, 이것이 모델에 문제를 발생시킨다고 할 수는 없을 것 같다.

3. 곡 간의 consign similiarity 계산

처음 10곡을 기준으로 cosine similiarity plot한 결과이다.



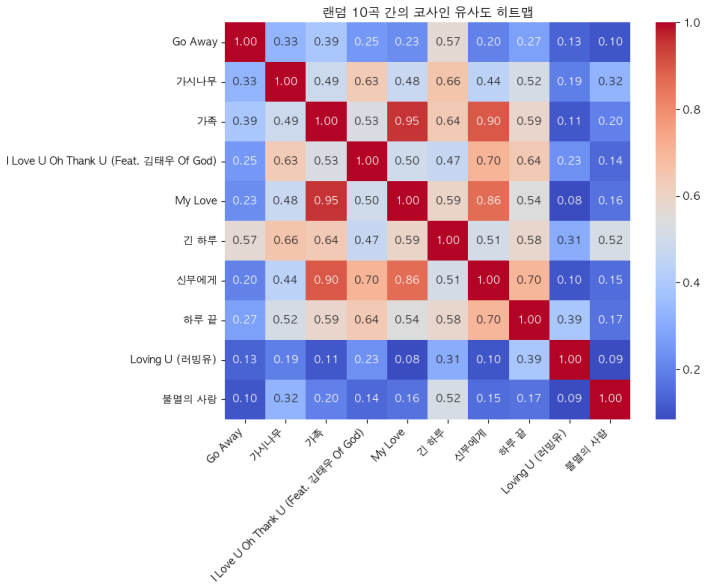
놀라운 유사도 결과를 확인할 수 있다.

1) 썸 - 금요일에 만나요
2) 야생화 - 눈, 코, 입
3) 한 여름밤의 꿀 - Mr.Chu

대체적으로 놀라울 정도의 유사도를 보여준다. 물론 해당 노래의 inference가 잘못된 경우에는 consine 유사도가 부정확해지는 모습도 확인된다.

특히 something(걸스데이)의 경우, 이별노래나 사랑 노래에 대체적으로 유사한 것으로 나타난다. 하지만 가사를 보면, 거짓이 들통난 남자친구에 대한 살짝의 증오가 담겨있다. 이를 보면 감정이 살짝 일반적이지 않은 경우, 모델의 감정분석에 오차가 있을 수 있음을 확인했다.

추가적으로 ranodm으로 10개를 sampling해서 유사도 heatmap를 그리면 아래와 같다.



여기서 불멸의 사랑(조성모)의 가사를 확인하면, 너는 나를 잊고 갔어도 나는 영원히 너를 사랑한다는 내용이다. 모델은 마비된을 50% 확률로 top1으로 inference한다. 여전히 해어진 연인을 계속 사랑한다는 내용은 흔한 내용은 아니라고 생각되고, heatmap에서 모든 노래와 푸른 계열을 띄고 있다.

4. 추천 시스템 preview

Recommendation Preview

아무 곡이나 3개를 선정하고, 해당 추천이 얼마나 시대적 거리감을 뛰어넘을 수 있는지를 확인하고, 유사성을 perceptual하게 확인한다.

눈물 (Feat. 유진 Of 더 씨야) - 리쌍

Rank	Song	Artist	Similarity
1	어인의 향기	씨야	0.9936
2	그녀의 연인에게... #Story 1	K2 김성면	0.9935
3	눈, 코, 입	태양	0.9917
4	No.1	보아 (BoA)	0.9901
5	미워도 다시 한번 (Original Ver.)	바이브	0.9896
6	미안해요	김건모	0.9892
7	심장이 없어	에이트	0.9889
8	술프도록 아름다운 ..	K2 김성면	0.9886
9	야생화	박효신	0.9881
10	벌써 일년	브라운 아이즈	0.9863

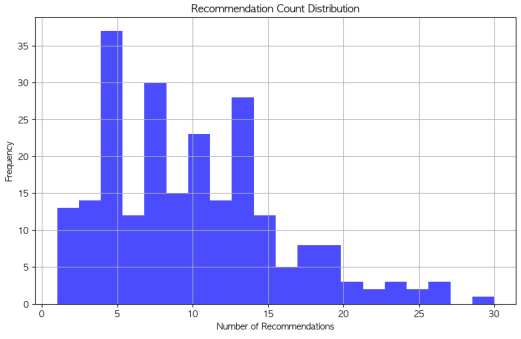
아로하 - 쿨 (COOL)

Rank	Song	Artist	Similarity
1	사랑의 서약	한동준	0.9732
2	Run To You	DJ DOC	0.8972
3	편지어	녹색지대	0.8796
4	너를 사랑해	한동준	0.8692
5	I Love U Oh Thank U (Feat. 김태우 Of God)	MC몽	0.7939
6	Mr. Chu (On Stage)	Apink (에이핑크)	0.7735
7	우린 제법 잘 어울려요	성시경	0.7665
8	멀어진 사람	임윤택 (Fin.K.L)	0.7575
9	너에게 원한건	노이즈	0.7072
10	Dreams Come True	S.E.S.	0.6983

내사랑 내곁에 (응답하라 1988 삽입곡) - 김현식

Rank	Song	Artist	Similarity
1	사랑...후에	신해성, 윤	0.9790
2	부디	윤종신	0.9781
3	추억은 사랑을 닮아	박효신	0.9770
4	사랑비	김태우	0.9707
5	꿈	이현우	0.9693
6	착한 사람	김민종	0.9672
7	슬픈 인연	O1SB	0.9665
8	오늘 같은 밤이면	박정운	0.9661
9	환눈	이루	0.9638
10	모노드라마 (With 유승우)	허각	0.9629

해당 예시들을 본다면 각 노래들에 대해서 다양한 시대의 가수들이 어우러져 있는 모습을 확인할 수 있다. 또한 해당 노래의 추천 수의 분포를 통해서 얼마나 고르게 추천이 되어 있는지를 확인하였을 때 아래와 같은 분포를 가진다.



5. consine similiarity 계산을 보완

해당 similiarity 값들을 기반으로 추천을 한 데이터를 보면 대부분의 데이터가 95이상인 것을 볼 수 있다. 이는 consine similiarity 계산에서 가중치를 부여함으로서 조금 더 곡 간의 유사도를 명확하게 할 수 있다고 생각하였다. 따라서 cosine 유사도 식을 아래와 같이 변경하였다.

(본래 식)

$$\text{Cosine Similarity}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

$$\mathbf{A}\cdot\mathbf{B} = \sum_{i=1}^{n} A_i B_i$$

$$\|\mathbf{A}\| = \sqrt{\sum_{i=1}^{n} A_i^2} \|\mathbf{B}\| = \sqrt{\sum_{i=1}^{n} B_i^2}$$

(아래와 같이 변경)

$$\text{Weighted Cosine Similarity}(\mathbf{A},\mathbf{B}) = \frac{\sum_{i=1}^{n} w_i A_i B_i}{\sqrt{\sum_{i=1}^{n} w_i A_i^2}\sqrt{\sum_{i=1}^{n} w_i B_i^2}}$$

특히 이때, weigh는 각 곡의 top 1의 감정에 가중치 2를 부여한다. 본래는 모두 동일한 가중치(1)을 부여한 것과 같다.
해당 consine 유사도로 노래 추천결과는 아래와 같다.

눈물 (Feat. 유진 Of 더 씨야) - 리쌍

Rank	Song	Artist	Similarity
1	여인의 향기	씨야	0.9982
2	그녀의 안인에게.. #Story 1	K2 김성빈	0.9982
3	눈, 코, 입	태양	0.9975
4	No.1	보아 (BoA)	0.9973
5	아워도 다시 한번 (Original Ver.)	바이브	0.9971
6	아안해요	김건모	0.9970
7	심장이 없어	에이트	0.9969
8	슬프도록 아름다운 ..	K2 김성빈	0.9968
9	아생화	박효신	0.9966
10	벌써 일년	브라운 아이즈	0.9961

아로하 - 쿨 (COOL)

Rank	Song	Artist	Similarity
1	사랑의 사막	한동준	0.9914
2	Run To You	DJ DOC	0.9648
3	환절아	녹색지대	0.9459
4	너를 사랑해	한동준	0.9389
5	Mr. Chu (On Stage)	Apink (에이핑크)	0.9073
6	I Love U Oh Thank U (Feat. 김태우 Of God)	MC몽	0.9062
7	우린 제법 잘 어울려요	성시경	0.8962
8	너에게 원한건	노어즈	0.8596
9	Dreams Come True	S.E.S.	0.8383
10	앨범한 사람	빌클 (Fin.K.L)	0.5260

내사랑 내곁에 (응답하라 1988 삽입곡) - 김현식

Rank	Song	Artist	Similarity
1	사랑...후에	신혜성, 린	0.9931
2	부디	윤종신	0.9928
3	추억은 사람을 달아	박효신	0.9923
4	사랑비	김태우	0.9905
5	꿈	이현우	0.9900
6	확한 사람	김민종	0.9892
7	슬픈 인연	015B	0.9890
8	오늘 같은 밤이면	박정운	0.9887
9	모노드라마 (With 유승우)	허각	0.9879
10	I Love You	포지션	0.9878

해당 matrix로 계산시 조금 더 명확한 추천이 가능할 것으로 예상되었으나, 본래 가중치 1로 진행한 추천과 순서만의 차이만 존재하였다. 따라서 기존 방식인 consine similiarity 계산도 적합함을 확인하였다.

5. 한계

1. 확실한 Multi-lingual의 부재

해당 모델은 klue/bert-base에 기반하여 fine-tunning되고 inference되었기 때문에 영어에 취약하다. 일례로 FANTASTIC BABY 가사를 통한 추천과 inference를 보면 아래와 같다.

FANTASTIC BABY - BIGBANG (빅뱅)

Rank	Song	Artist	Similarity
1	Abracadabra	브라운아이드걸스	0.9095
2	아주 가끔은	신해철	0.8715
3	죽을 만큼 아파서 (Feat. 멜로우)	MC몽	0.8706
4	사랑과 전쟁 (Narr. 하하)	다비치	0.8458
5	겁쟁이	버즈	0.8431
6	바람났어 (Feat. 박봄)	GG (박명수 & G-Dragon)	0.8222
7	내가 제일 잘 나가	2NE1	0.8218
8	? (물음표) (Feat. 최자 Of 다이나믹 듀오, Zion.T)	프라이머리	0.8148
9	Again & Again	2PM	0.8055
10	길	god	0.7620

FANTASTIC BABY

Rank	Prediction	Probability
1	안달하는	0.0824
2	분노	0.0776
3	환멸을 느끼는	0.0717
4	구역질 나는	0.0548
5	고립된(당황한)	0.0522
6	마비된	0.0399
7	혼란스러운	0.0374
8	당혹스러운	0.0335
9	짜증내는	0.0290
10	억울한	0.0284

사실 이는 가사의 감정 분석의 정확성의 문제라기 보다는, 가사를 통한 의미 전달보다 audio의 가중치가 더 큰 음악의 경우 가사를 기반으로 추천하는 것의 한계가 있을 것이라는 해석이 더 정확하다.

2. 확실한 평가지표의 부재

감정 분석 자체의 benchmark뿐만 아니라, 해당 노래 추천 알고리즘의 평가도 불가능하기 때문에 시장에 테스트를 통해서만 소비자의 반응을 알 수 있다는 한계가 존재한다. 이를 해결하기 위해서는 기존에 만들어진 playlist를 기반으로 평가지표를 마련하거나, 해당 playlist를 기반으로 학습을 진행한다면 조금 더 우수한 모델이 만들어질 것으로 예상된다.

6. 정리

해당 프로젝트를 통해서 먼저 가사 데이터셋이 존재하는 영어 PLM을 기반으로 baseline을 구축하고, 학습을 위한 qlora의 역할을 확인하였습니다.

그 후 감정분석 모델 자체의 architecture에 집중하여서 klue/bert-base를 기반으로 tunning을 진행하였습니다. 총 2가지 모델을 학습시켰고, 이미 존재하는 hugging-face hun3359/klue-bert-base-sentiment 모델과 상호비교를 통해 모델의 성능을 향상시켰습니다.

특히, architecture2는 모델의 짧은 text의 학습 데이터를 긴 text의 가사에 적용하기 위해서 attention mask를 활용한 layer를 classifier에 추가하였고, 768차원 bert output의 특성을 최대한 살리기 위해서 노력하였습니다. 그 결과 proposed 모델을 감정의 통일성에 강점을 가졌고, 이는 감정 자체의 유사도까지 동시에 학습할 수 있는 모델임을 확인할 수 있었다.

해당 모델(archictecture2)를 바탕으로 향후 연구를 진행하였다.
가사를 멜론 음원 사이트에서 다양한 시대가 포함된 것으로 crowling하였고, 해당 노래들의 songid를 바탕으로 가사도 chrome driver를 활용하여 crowling하였다.
이를 바탕으로 architecture2모델로 감정분석을 inference하였다.
Inference결과를 바탕으로 60차원 vector의 consine 유사도를 계산하여 각 곡 마다 10곡을 추천하는 시스템을 구성하였다.

그 결과 본래 목적이였던, 감정이 비슷하지만 시간적 거리가 존재하는 두 곡 간의 유사도를 얻을 수 있었고, 이를 기존의 사용자 로그 기반 추천에 적용한다면 시너지를 확보할 수 있을 것으로 예측된다.

References

https://www.eliftech.com/insights/all-you-need-to-know-about-a-music-recommendation-system-with-a-step-by-step-guide-to-creating-it/

https://www.melon.com/m6/landing/djplayList.htm?type=djc&plylstSeq=410441907#params%5BplylstSeq%5D=410441907&po=pageObj&startIndex=101

https://developer.chrome.com/docs/chromedriver/get-started?hl=ko

https://huggingface.co/hun3359/klue-bert-base-sentiment
https://huggingface.co/klue/bert-base