MobileBERT 활용한 YouTube 뮤직 비디오 댓글 감성 분석 프로젝트

+ 프로젝트 개요

이 프로젝트는 K-pop을 대표하는 명곡 소녀시대의 '다시 만난 세계 (Into The New World)' 뮤직비디오에 달린 영어 댓글을 수집하고, 이를 분석하여 사람들의 감정 반응을 MobileBERT 기반 감성 분석 모델로 파악하는 것을 목표로 합니다. 선정한 곡은 한국 대중문화에서 상징적인 의미를 가지며, 글로벌 K-pop 팬들 사이에서도 지속적으로 언급되는 곡입니다. 본 프로젝트는 전 세계 사람들이 해당 곡에 대해 어떤 감정을 표현하고 있는지를 데이터 기반으로 분석하고 시각화 함으로써, 문화적·정서적 반응의 흐름을 이해하고자 합니다.

+ 개발 동기

기존 공공도서관을 검색하는 시스템은 지역별로 검색하였을 때에 목록으로만 정보를 확인할 수 밖에 없었고, 직관적로 위치를 지도상에 표시를 해주지 않기 때문에 지도를 추가하면 좋겠다는 생각을 했습니다

또한, 기존 검색 시스템에서는 도서관의 이름, 주소, 영업일에 대해서만 정보를 확인할 수 있었지만, 저희는 기존 정보도 제공하면서 열람좌석, 보유도서수, 대출가능권수, 일수와 같은 도서관 방문자가 궁금해할만한 자세한 정보까지 제공하기 위하여 개발을 하게 되었습니다.

+ 데이터 처리 과정

1. 원시 데이터 수집 (JSONL)

2-2. JSONL 데이터 형식을 변환한 CSV 형식의 2,000개의 영어 댓글 수동 데이터 라벨링

CSV 원시 데이터 형식 예시

필드명	설명
comment	유튜브 댓글 내용 (영어로 된 댓글 텍스트)
label	수동 감성 라벨 (0: 부정, 1: 중립, 2: 긍정)

예시 (5개 댓글)

comment (댓글 내용)	label (감성 라벨)
This is not a "song", This is an "anthem"	2
I love this song!	2
The video quality is poor.	0
Not my favorite but still good.	1
The choreography is amazing!	2
... (총 2,000개의 댓글과 라벨)	...

2. 데이터 수동 라벨링 (CSV)

2-2. JSONL 데이터 형식을 변환한 CSV 형식의 2,000개의 영어 댓글 수동 데이터 라벨링

CSV 원시 데이터 형식 예시

필드명	설명
comment	유튜브 댓글 내용 (영어로 된 댓글 텍스트)
label	수동 감성 라벨 (0: 부정, 1: 중립, 2: 긍정)

예시 (5개 댓글)

comment (댓글 내용)	label (감성 라벨)
This is not a "song", This is an "anthem"	2
I love this song!	2
The video quality is poor.	0
Not my favorite but still good.	1
The choreography is amazing!	2
... (총 2,000개의 댓글과 라벨)	...

3. 데이터 토큰화

3-1. 토큰화 결과 샘플

```
### 토큰화 결과 샘플 ###

1번째 데이터
데이터 : This is not a "song" , This is an "anthem"
토큰 : [101, 2023, 2003, 2025, 1037, 1000, 2299, 1000, 1010, 2023, 2003, 2019, 1000, 11971, 1000, 102, 0, 0, 0, ..., 0]
어텐션 마스크 : [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, ..., 0]

2번째 데이터
데이터 : 657K to 57M fighting sones fighting 💪 🖤 💙 fighting 💪 💙 🖤
토큰 : [101, 3515, 2581, 2243, 2000, 5401, 2213, 3554, 2365, 2229, 3554, 100, 100, 100, 3554, 100, 100, 100, 102, 0, 0, 0, ..., 0]
어텐션 마스크 : [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, ..., 0]

3번째 데이터
데이터 : Never too old!
토큰 : [101, 2196, 2205, 2214, 999, 102, 0, 0, 0, ..., 0]
어텐션 마스크 : [1, 1, 1, 1, 1, 1, 0, 0, 0, ..., 0]
```

4. 데이터 학습 결과

3-2. 학습 결과

```
Training Epoch 1: 100%|██████████| 200/200 [00:17<00:00, 11.16it/s, loss=0.619]
Evaluaction Train Epoch 1: 100%|██████████| 200/200 [00:04<00:00, 47.18it/s]
Evaluation Validation Epoch 1: 100%|██████████| 50/50 [00:01<00:00, 46.64it/s]
Training Epoch 2: 100%|██████████| 200/200 [00:17<00:00, 11.14it/s, loss=0.00719]
Evaluaction Train Epoch 2: 100%|██████████| 200/200 [00:04<00:00, 46.58it/s]
Evaluation Validation Epoch 2: 100%|██████████| 50/50 [00:01<00:00, 47.58it/s]
Training Epoch 3: 100%|██████████| 200/200 [00:17<00:00, 11.39it/s, loss=0.00782]
Evaluaction Train Epoch 3: 100%|██████████| 200/200 [00:04<00:00, 46.49it/s]
Evaluation Validation Epoch 3: 100%|██████████| 50/50 [00:01<00:00, 43.09it/s]
Training Epoch 4: 100%|██████████| 200/200 [00:28<00:00, 7.07it/s, loss=0.741]
Evaluaction Train Epoch 4: 100%|██████████| 200/200 [00:07<00:00, 26.52it/s]
Evaluation Validation Epoch 4: 100%|██████████| 50/50 [00:01<00:00, 25.91it/s]
Epoch 1: Train loss: 73355.7421, Trian Accuracy: 0.8775, Validation Accuracy: 0.8525
Epoch 2: Train loss: 0.7837, Trian Accuracy: 0.8806, Validation Accuracy: 0.8525
Epoch 3: Train loss: 0.4041, Trian Accuracy: 0.9006, Validation Accuracy: 0.7625
Epoch 4: Train loss: 0.3994, Trian Accuracy: 0.9094, Validation Accuracy: 0.8525

### 모델 저장 ###
모델 저장 완료

종료 코드 0(으)로 완료된 프로세스
```

+ 프로젝트 결론 및 소감

3-3. 학습 결과 분석

학습 및 검증 성능

에폭(Epoch)	학습 손실(Train Loss)	학습 정확도(Train Accuracy)	검증 정확도(Validation Accuracy)
1	0.619 (단, 73355.7421라는 비정상적 값이 기록됨)	87.75%	85.25%
2	0.00719	88.06%	85.25%
3	0.00782	90.06%	76.25%
4	0.741	90.94%	85.25%

주요 관찰점

손실 및 정확도 변화

- 학습 정확도는 에폭이 진행될수록 꾸준히 상승하여 90% 이상에 도달했습니다.
- 검증 정확도는 대체로 85% 근처에서 안정적이었으나, 3번째 에폭에서는 76.25%로 급락하여 불안정한 모습을 보였습니다.
- 학습 손실은 2~3번째 에폭에서 매우 낮은 값을 기록하였고, 1번째 에폭에서는 비정상적으로 높은 값이 관찰되어 손실 계산이나 로깅에 오류가 있을 가능성이 있습니다.
- **과적합 가능성**
  - 학습 정확도가 상승하는 동안 검증 정확도가 불안정하거나 떨어지는 현상은 모델이 학습 데이터에 과적합되고 있을 가능성을 시사합니다.
- **로깅 문제**
  - 1번째 에폭의 학습 손실 값(73355.7421)은 비정상적으로 높아 로그 오류나 계산 오류로 판단됩니다.
  - 손실 값의 일관성이 부족하여 정확한 학습 모니터링에 어려움이 있습니다.
- **성능 및 속도**
  - 학습 속도는 초당 약 7 에서 11 iter 이며 평가 속도는 약 25 에서 47 iter 이므로 하드웨어 사양 대비 적절한 수준입니다.

이번 프로젝트를 통해 데이터 수집, 데이터 정제, 데이터 라벨링, 데이터 분석을 혼자서 직접 배우며 어려움을 극복해나가는 과정이였습니다. 특히, 단 '2,000건'의 수작업 라벨링 데이터만으로도 '약 87%' 이상의 정확도를 달성한 것은 매우 인상 깊은 결과라고 느끼고 실력이 좋지는 않지만 나름 괜찮은 결과를 도출했다고 생각을 합니다. 여러 과정중에서 데이터들을 수집하는 것도 그렇고 '2,000건'이라는 데이터를 직접 라벨리을 하는 것은 결코 쉬운 과정은 아니였고 영어로 작성된 댓글들을 이해하여 직접 라벨리을 분류하여 작성한다는 것이 쉽지가 않았습니다. 데이터라는 것을 다룬다는 것이 결코 코드를 작성하는 것만 있는 것이 아니라 프로젝트에 대한 목적을 선정하고 그 목적을 달성하기 위한 과정이 중요하다는 것을 직접 경험했다는 것이 중요했다고 생각합니다. 프로젝트라는 것을 이번 기회를 통해 처음 단독으로 시작하여 직접 주제의 선정부터 보고서 작성이라는 다양한 과정의 마무리까지 마치며 한계를 극복하는 좋은 경험을 했다고 생각하며 마무리를 했습니다.