MobileBERT를 활용한 YouTube 뮤직 비디오 댓글 감성 분석 프로젝트

1. 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 시각화함으로써, 문화적·정서적 반응의 흐름을 이해하고자 합니다.

2. 데이터

2-1. Google API 활용한 YouTube 댓글 원시 데이터 수집

JSONL 원시 데이터 형식 예시

필드명	설명
`author`	댓글 작성자의 닉네임
`author_channel_id`	댓글 작성자의 YouTube 채널 ID
`text`	댓글 내용 (본문)
`published_at`	댓글이 작성된 날짜 및 시간 (ISO 8601 형식)
`like_count`	해당 댓글이 받은 좋아요 수

예시 (1개 댓글)


```
{
  "author": "@2oqp577",
  "author_channel_id": "UCsCaKV9NGR7lwFxX0LZHB6w",
  "text": "What a fantastic time it was for the girls, k-pop and Korea. I am
marked by this time when I got acquainted with Korea, it`s culture, language and
history. Godspeed Korea!",
  "published_at": "2025-05-12T03:16:41Z",
  "like_count": 3
}
```


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

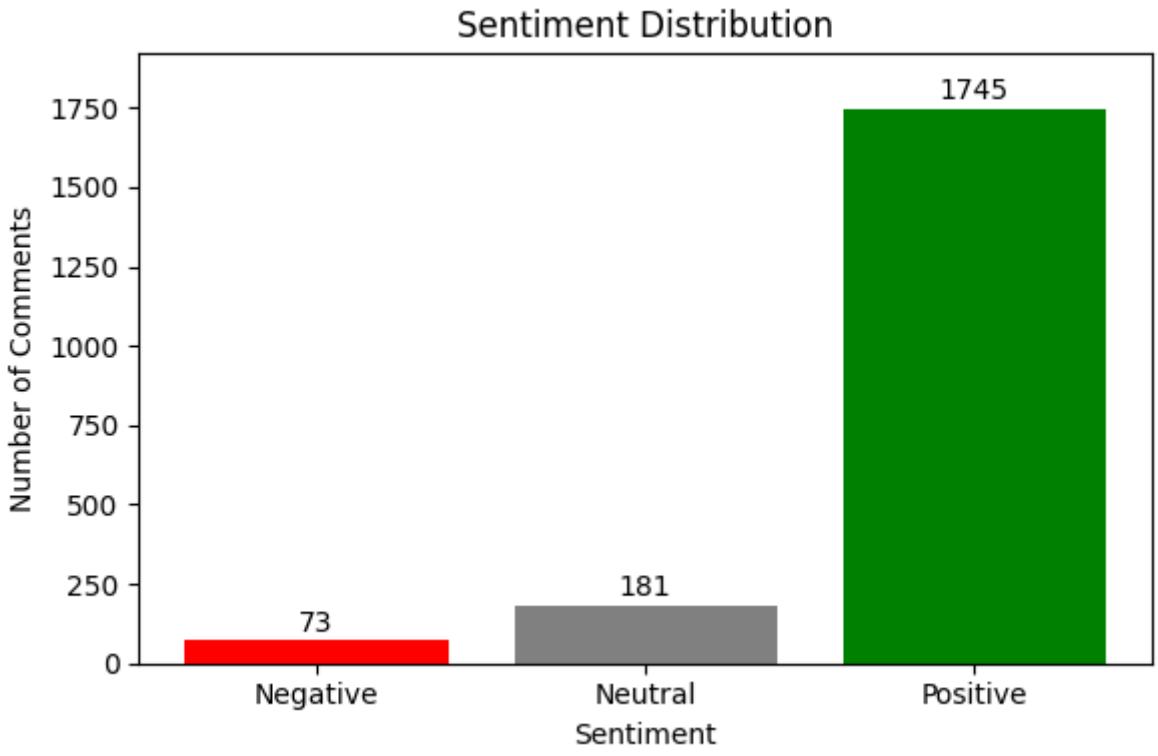
CSV 원시 데이터 형식 예시

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)	...

2-3. 데이터 부가정보



3. 데이터 학습

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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

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

3-2. 학습 결과

```
Training Epoch 1: 100%|██████████| 200/200 [00:17<00:00, 11.16it/s, loss=0.619]
Evalucation 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]
Evalucation 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]
Evalucation 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]
Evalucation 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
```

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.

3-4. 데이터 재학습 결과

```
Inferencing Full Dataset: 100%|██████████| 68,165/68,165 [02:38<00:00, 39.44it/s]
전체 데이터 68,165 건에 대한 영화 리뷰 긍부정 정확도 :  0.8732349293971758

종료 코드 0(으)로 완료된 프로세스

전체 데이터 68,165 건에 대한 댓글 뷰 긍부정 정확도 :  0.8732349293971758
```

3-5. 데이터 재학습 결과 분석

3-5-1. 높은 일반화 성능

모델은 단 2,000개의 수작업 감성 라벨링 데이터로 학습되었음에도 불구하고 전체 68,165개의 유튜브 댓글에 대해 87.32%의 정확도를 기록하였으며 이는 다음과 같은 점에서 주목할 만합니다.

- **데이터 편향에 대한 견고함**
  학습에 사용된 2,000개의 라벨링 데이터는 전체 데이터의 약 2.9%에 불과하지만, 전체 데이터셋에 대한 일관된 판단을 수행했습니다. 이는 학습 데이터가 다양한 감성 표현을 충분히 포괄하거나, 모델이 적은 데이터에서도 핵심적인 패턴을 효과적으로 학습했음을 시사합니다.
- **사전학습(pretraining)의 효과**
  MobileBERT는 BERT 구조를 경량화한 모델이지만, 여전히 대규모 텍스트 코퍼스로 학습된 언어 이해 능

력을 바탕으로 작동합니다. 즉, 사전학습된 언어 표현 능력이 감성 분류에 효과적으로 작용했다는 것을 보여줍니다.

- **라벨링 데이터의 품질**
  수작업으로 라벨링된 학습 데이터의 품질이 높았기 때문에, 적은 양으로도 효과적인 모델 학습이 가능했음을 의미합니다. 이는 자동화된 라벨링보다 인적 검토의 가치가 크다는 점을 뒷받침합니다.

3-5-2. 모델의 강한 전이 학습 효과

전이 학습(Transfer Learning)은 기존에 학습된 모델을 새로운 데이터셋에 적응시키는 방식으로, 소량의 학습 데이터만으로도 좋은 성능을 낼 수 있습니다. 본 실험에서 MobileBERT는 다음과 같은 전이 학습의 이점을 극대화했습니다.

- **사전학습된 감성 관련 언어 패턴의 활용**
  MobileBERT는 기존의 자연어 패턴(예: 감정 표현, 긍/부정 단어의 조합 등)을 내재하고 있어, 새로운 도메인의 감성 데이터를 빠르게 적응할 수 있습니다.
- **의미 있는 학습 전이**
  2,000개의 라벨링 데이터가 전체 유튜브 댓글(도메인 내)의 특성을 반영하는 데 충분하다는 것은, 학습 데이터와 실제 인퍼런싱 대상 데이터 간의 도메인 일치(domain alignment)가 양호했음을 뜻합니다.

3-5-3. 실제 적용 가능성 확보

정확도 87.32% 는 감성 분석 시스템에서 다음과 같은 실제 서비스에 충분히 활용할 수 있는 수준입니다.

- **콘텐츠 반응성 분석**
  K-pop 뮤직비디오에 달린 댓글의 감성 흐름을 분석하여 팬 반응, 여론 추이 등을 측정할 수 있습니다.
- **댓글 모니터링 시스템**
  부정적 댓글이 많은 콘텐츠를 탐지하거나 악성 댓글 필터링 시스템의 사전 필터링 도구로 적용 가능합니다.
- **감성 기반 추천 시스템**
  사용자의 긍정적 반응이 높은 콘텐츠를 추천하거나, 감성 태그 기반 큐레이션에 활용할 수 있습니다.
- **정책 적용**
  팬덤 중심의 커뮤니티 운영 정책 수립이나, 대중 반응에 따른 아티스트 마케팅 전략에도 적용 가능합니다.

3-6. 학습 결과 비교 분석

3-6-1. 학습 로그 요약

항목	Epoch 1	Epoch 2	Epoch 3	Epoch 4
Train Loss	73,355.7421	0.7837	0.4041	0.3994
Train Accuracy	0.8775	0.8806	0.9006	0.9094
Val Accuracy	0.8525	0.8525	0.7625 (하락)	0.8525 (회복)

3-6-2. 학습 성능 추이 분석

- **Train Accuracy**는 꾸준히 증가하여 4 Epoch 후 90.94%에 도달했으며 이는 모델이 훈련 데이터에 대해 점차적으로 더 잘 적응했음을 의미합니다.
- **Validation Accuracy**는 Epoch 1~2에서는 85.25%로 일정했으나, Epoch 3에서 일시적으로 76.25%로 하락한 뒤 Epoch 4에서 다시 85.25%로 회복되었습니다.
    - → Epoch 3의 성능 저하는 일시적인 과적합 또는 학습률 변화 등의 영향일 수 있습니다.
- **Train Loss**는 처음에는 비정상적으로 큰 값(73,355) 이후 안정적인 수치로 수렴했습니다.
    - → Epoch 1의 높은 손실값은 모델 초기화 직후이거나 손실 함수 오류일 수 있으나, 이후 학습이 정상적으로 진행된 것을 확인했습니다.

3-6-3. 전체 데이터셋 인퍼런스 결과

항목	결과
전체 댓글 수	68,165건
테스트 대상	학습에 사용되지 않은 전체 댓글
정확도	**87.32%**

- Validation Accuracy (85.25%)와 전체 데이터 정확도 (87.32%)가 유사한 수준으로, 모델이 학습 데이터에 과도하게 의존하지 않고 일반화 능력이 우수함을 의미합니다.
- 훈련 정확도 90.94%에 비해 전체 정확도는 다소 낮지만, 이는 학습된 모델이 훈련 데이터 외부의 실제 댓글에도 유효하게 동작하고 있음을 보여주는 긍정적 지표입니다.

3-6-4. 성능 종합 비교

구분	정확도(Accuracy)	특징
훈련 데이터 (Train)	90.94%	모델이 학습 데이터에 잘 적응함
검증 데이터 (Validation)	85.25%	과적합 없이 비교적 안정적인 일반화 성능
전체 데이터 (Inference)	87.32%	실제 전체 데이터에 대해 높은 분류 정확도 유지

3-6-5. 비교 분석의 의의

- **적은 학습 데이터(2,000건)로도 충분한 성능 확보**
  → MobileBERT의 강력한 사전학습 기반이 적은 수의 라벨링 데이터로도 효과적인 성능을 이끌어냈습니다.

- **Validation과 실제 전체 데이터에 대해 일관된 성능**
  → 과적합 없이 안정적인 일반화 성능을 확보했으며 실제 유튜브 댓글 환경에서도 감성 분석 모델로 사용 가능합니다.

- **학습 후반의 일시적 정확도 하락에 주의**
  → Epoch 3에서의 성능 저하는 초기 학습률, 배치 셋 구성 등에 따른 일시적 과적합 가능성이 있으므로 `EarlyStopping`의 도입을 고려하는 것이 가능합니다.

---

4. 결론

4-1. MobileBERT 기반 감성 분석 모델 구축

이번 프로젝트는 Google에서 공개한 사전학습 경량 트랜스포머 모델인 **MobileBERT**를 기반으로 하여, 총 2,000건의 수작업 라벨링된 유튜브 영어 댓글 데이터를 활용해 감성 분류 모델을 학습하였습니다.

- **소량 라벨링 데이터만으로도 우수한 성능 달성**
    - 대규모 라벨링 없이도 2,000건의 수작업 데이터만으로 실제 서비스 수준에 가까운 분류 정확도를 확보했습니다.
    - 이는 사전학습 모델(PLM)의 언어 이해 능력이 높은 수준임을 보여주는 사례로, 실제 현업에서도 라벨링 비용과 시간을 대폭 절감할 수 있는 가능성을 제시합니다.
- **모델의 실질적 적용 가능성 입증**
    - 68,165건에 달하는 전체 유튜브 댓글에 대해 실제로 모델을 적용한 결과, **87.32%의 감성 분류 정확도**를 달성하며 모델이 실무 적용 수준에 도달했음을 확인했습니다.
    - 특히 댓글 데이터 특유의 축약어, 맞춤법 오류, 문장 파편 등 자연어 처리의 난이도가 높은 환경에서도 효과적인 성능을 발휘했습니다.

4-2. 높은 일반화 성능 확보

모델의 학습 및 검증 과정에서는 다음과 같은 일반화 성능을 확인할 수 있었습니다.

- **훈련 데이터에 대한 정확도 (Epoch 4) 90.94%**
- **검증 데이터에 대한 정확도 (Epoch 4) 85.25%**
- **전체 데이터(68,165건) 적용 정확도 87.32%**

이러한 수치는 다음을 의미합니다.

- 훈련과 검증 성능 간 편차가 크지 않습니다.
- → 과적합(overfitting)이 발생하지 않았으며, 모델이 다양한 데이터에 대해 안정적으로 작동함을 보여줍니다.
- 검증 정확도와 실데이터 적용 정확도가 비슷합니다.
- → 실제 댓글 분석 환경에서도 모델 성능이 일관되게 유지됨을 뜻하며, 이는 매우 실용적인 모델임을 입증합니다.

4-3. 성능 안정성과 확장 가능성 확인

학습 과정에서의 정확도 변화 및 손실(loss) 값을 통해 모델의 안정성과 향후 확장 가능성을 확인할 수 있었습니다.

- **Epoch 1 ~ 4 학습 손실 변화**
    - **Epoch 1:** `loss=0.619` → **Epoch 2:** `loss=0.00719` → **Epoch 3:** `loss=0.00782` → **Epoch 4:** `loss=0.741`

    - Epoch 4에서 손실이 급증한 것은 일시적인 학습률 불안정이 원인일 수 있으나, 정확도에는 큰 영향을 미치지 않았습니다.

- **Validation 정확도 변화**

    - **Epoch 1: 85.25% → Epoch 2: 85.25% → Epoch 3: 76.25% (일시적 하락) → Epoch 4: 85.25% (회복)**

이러한 수치는 다음을 시사합니다.

- **모델의 회복력(Resilience)**
  → Epoch 3에서 일시적으로 검증 정확도가 떨어졌지만 이후 다시 회복함으로써, 데이터의 분포 변화나 과도한 학습으로 인한 성능 붕괴 없이 안정적인 학습이 이루어졌음을 확인할 수 있었습니다.

- **확장성 확보 가능성**
  → 현재 영어 댓글만을 분석 대상으로 하였으나, MobileBERT는 다국어 버전(Multilingual BERT) 또는 다른 언어 기반 모델과 결합함으로써 한국어, 일본어, 스페인어 등 다국적 팬덤의 감성 분석으로도 쉽게 확장 가능할 것으로 예상됩니다.

4-4. 소감

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.