BERT 모델

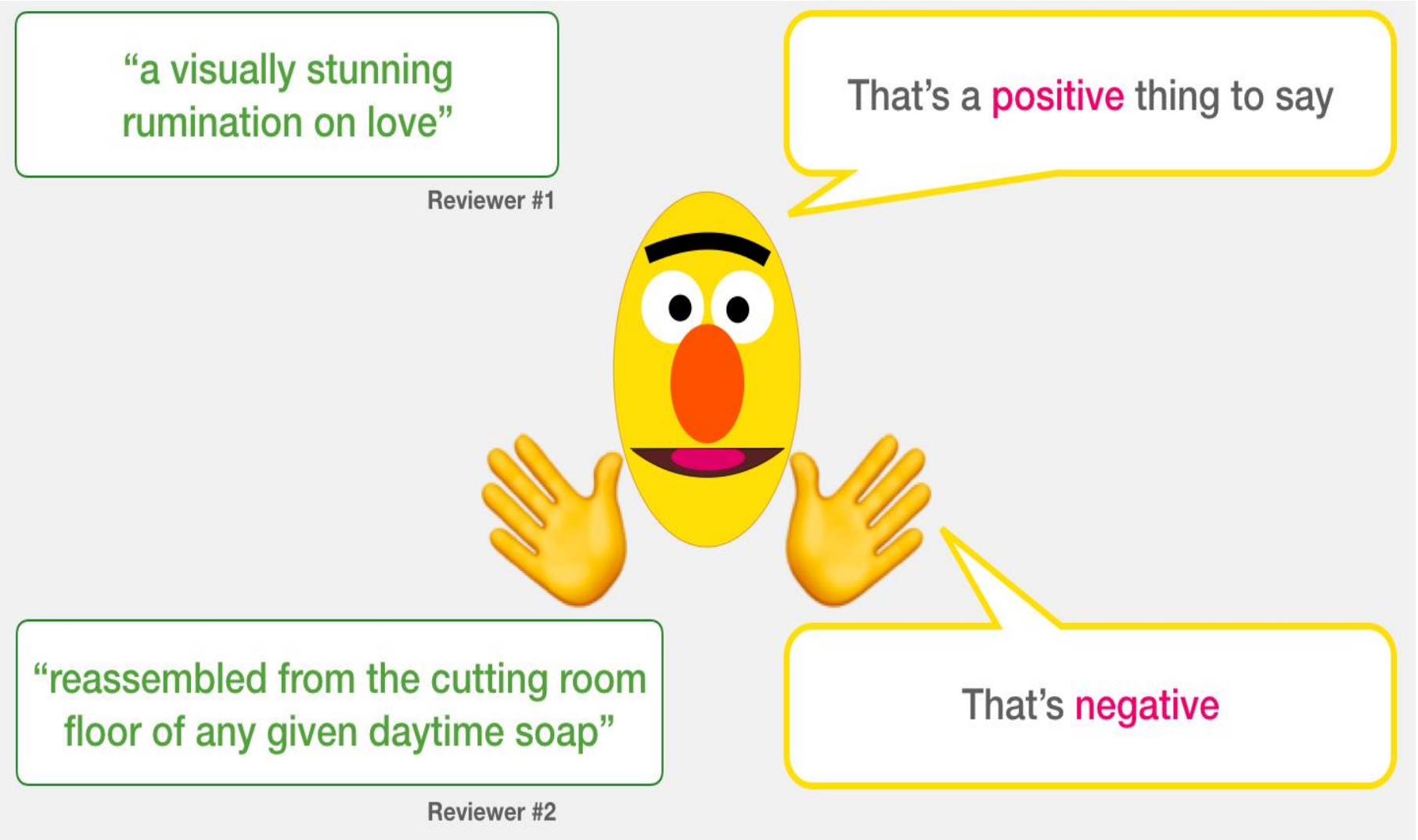
트랜스포머 양방향 인코더

Dr. Rhee
Feb 2020

BERT 응용

BERT 사용 가이드

- 영화 리뷰 감성 분석



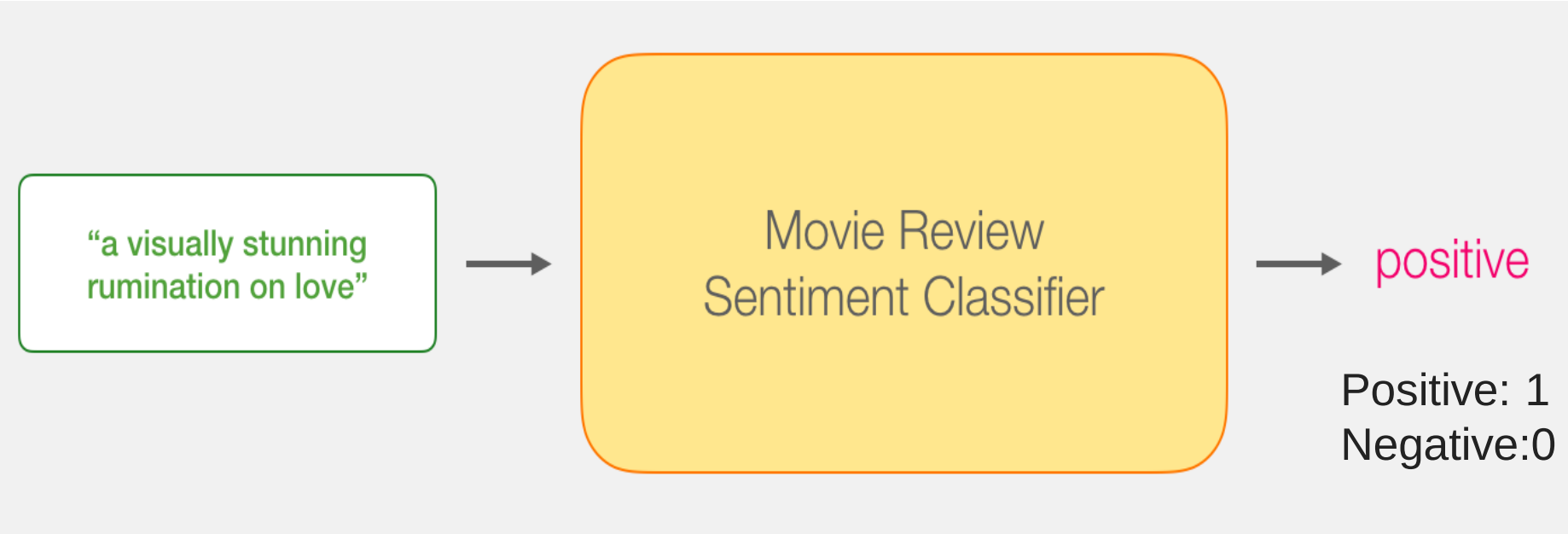
BERT 사용 가이드

- 데이터셋 SST2 (standford)

sentence	label
a stirring , funny and finally transporting re imagining of beauty and the beast and 1930s horror films	1
apparently reassembled from the cutting room floor of any given daytime soap	0
they presume their audience won't sit still for a sociology lesson	0
this is a visually stunning rumination on love , memory , history and the war between art and commerce	1
jonathan parker 's bartleby should have been the be all end all of the mod ern office anomie films	1

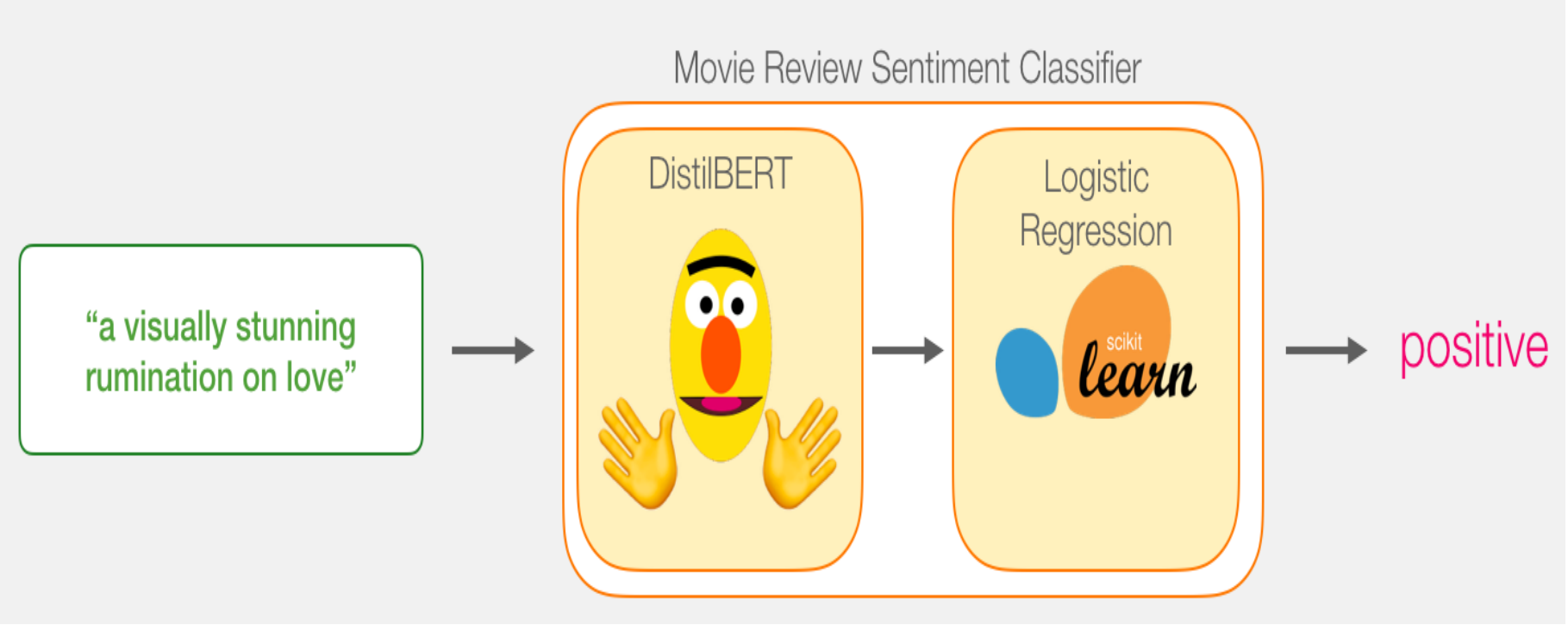
BERT 사용 가이드

- 모델: 문장 감성 분류.



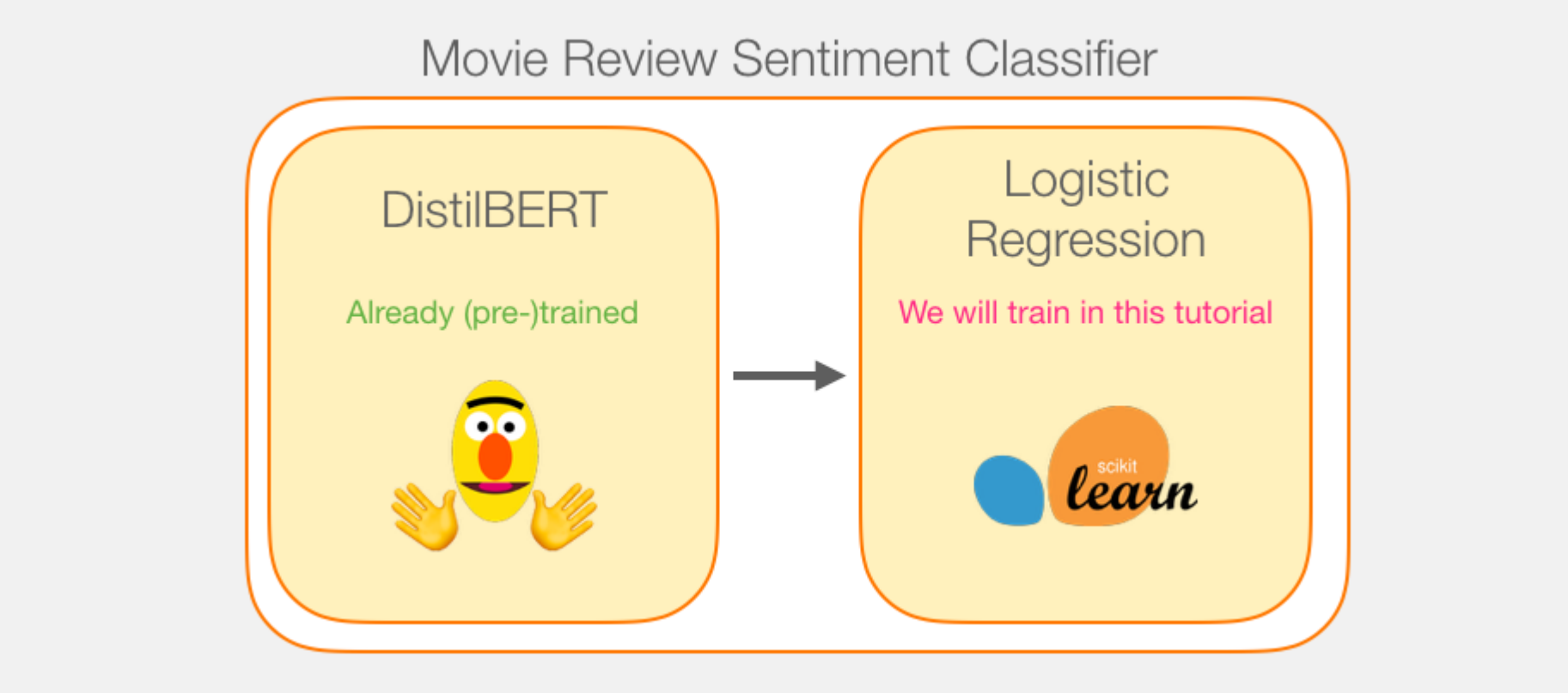
BERT 사용 가이드

- HuggingFace.가 개발한 소형 버전의 BERT인 DistilBERT를 사용한다. 두번째 모델인 로지스틱 회귀는 DistilBERT의 출력을 받아서 문장이 0 또는 1인지 분류한다.



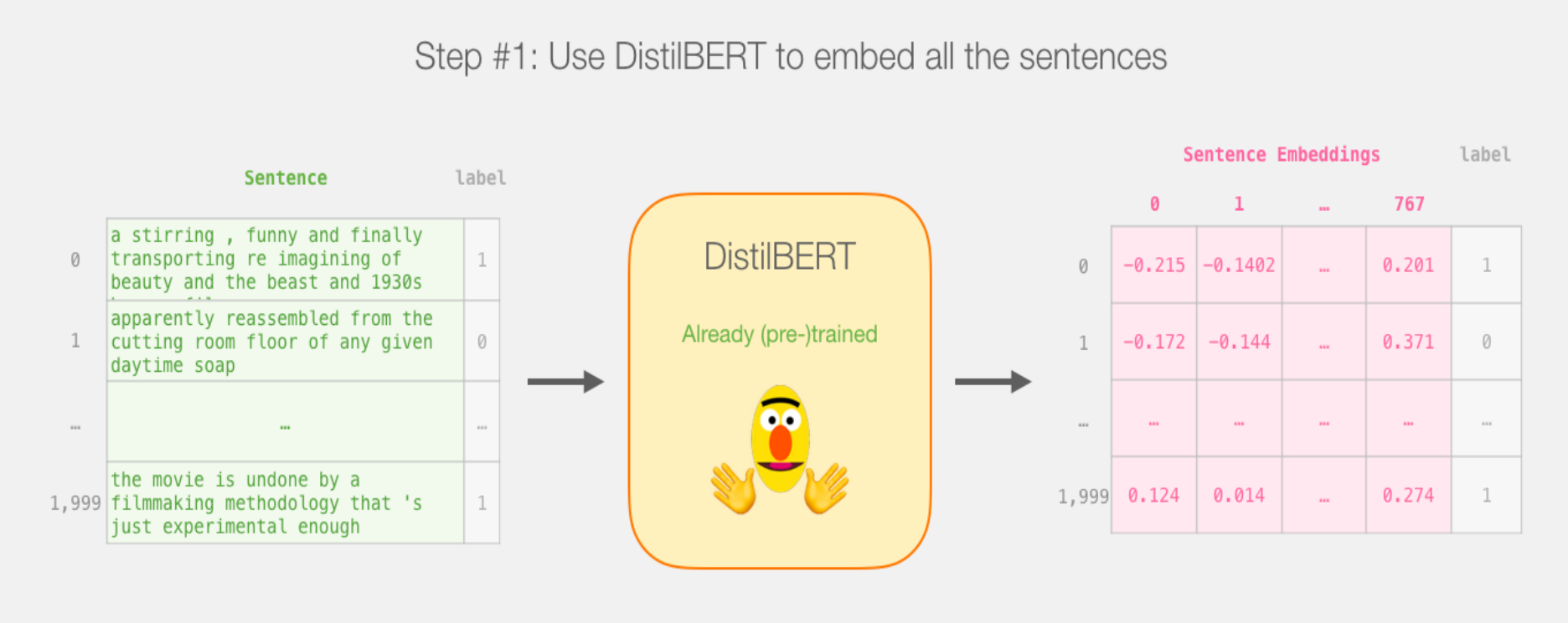
모델 학습

- DistitlBERT는 사전학습된 모델이며, 로지스틱 회귀만 학습한ㄷ.



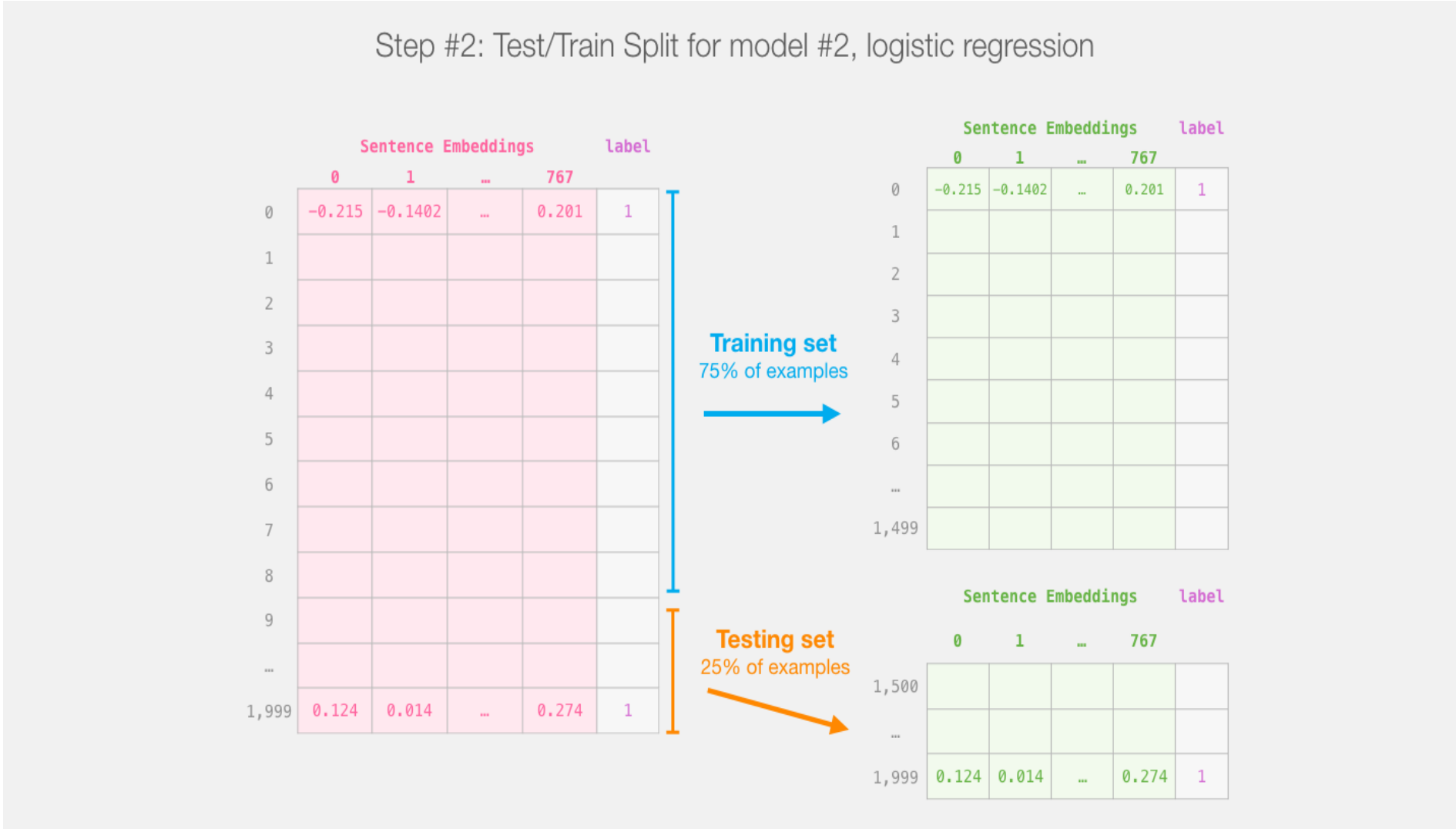
튜토리얼 개요

- 사전학습된 DistilBERT을 이용해 20,000개의 문장에 대한 sentence embedding을 얻는다.



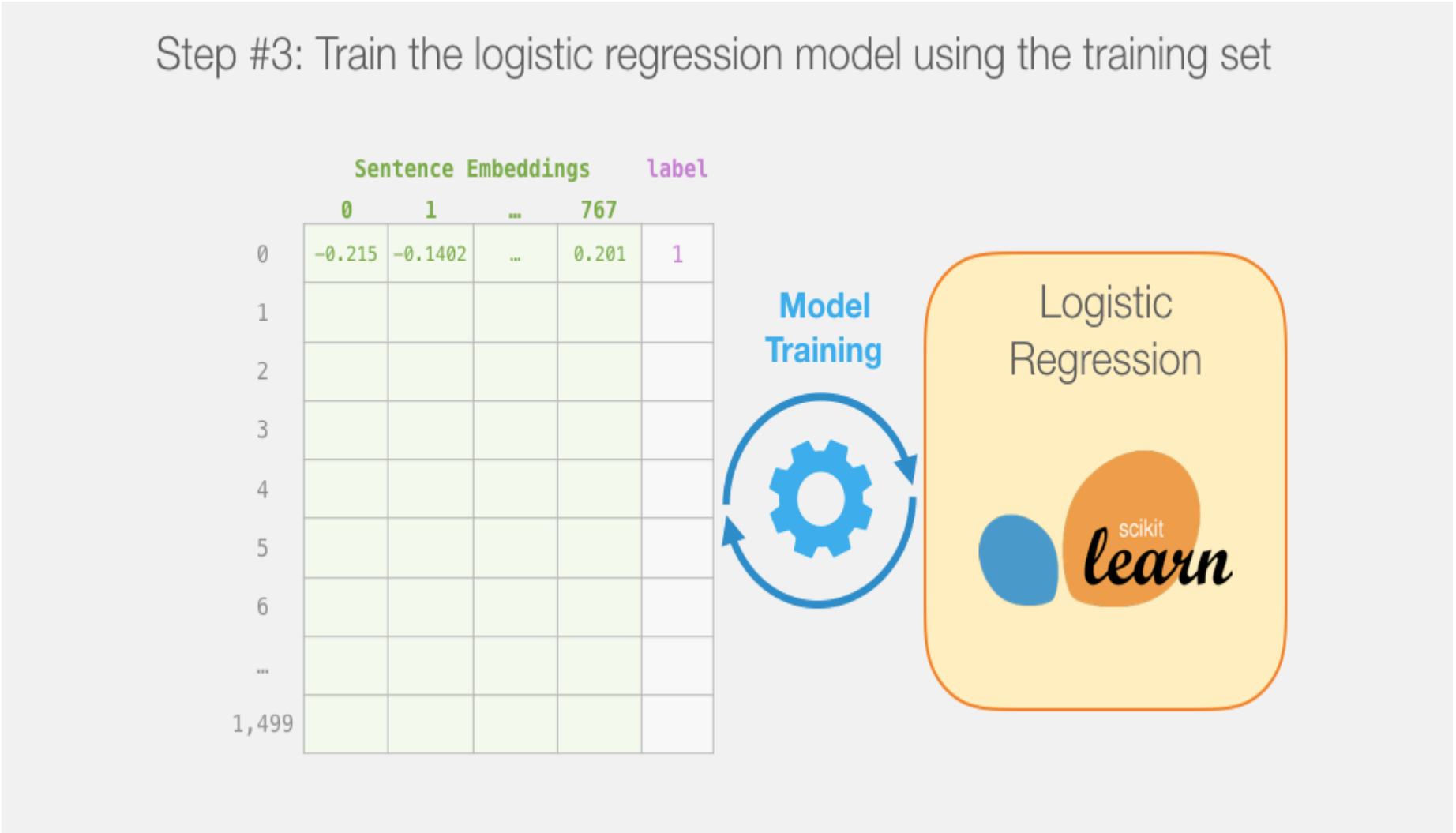
튜토리얼 개요

- 일반적인 훈련/테스트셋 분리를 한다. (DistlBERT는 이제부터 전혀 건드리지 않는다.)



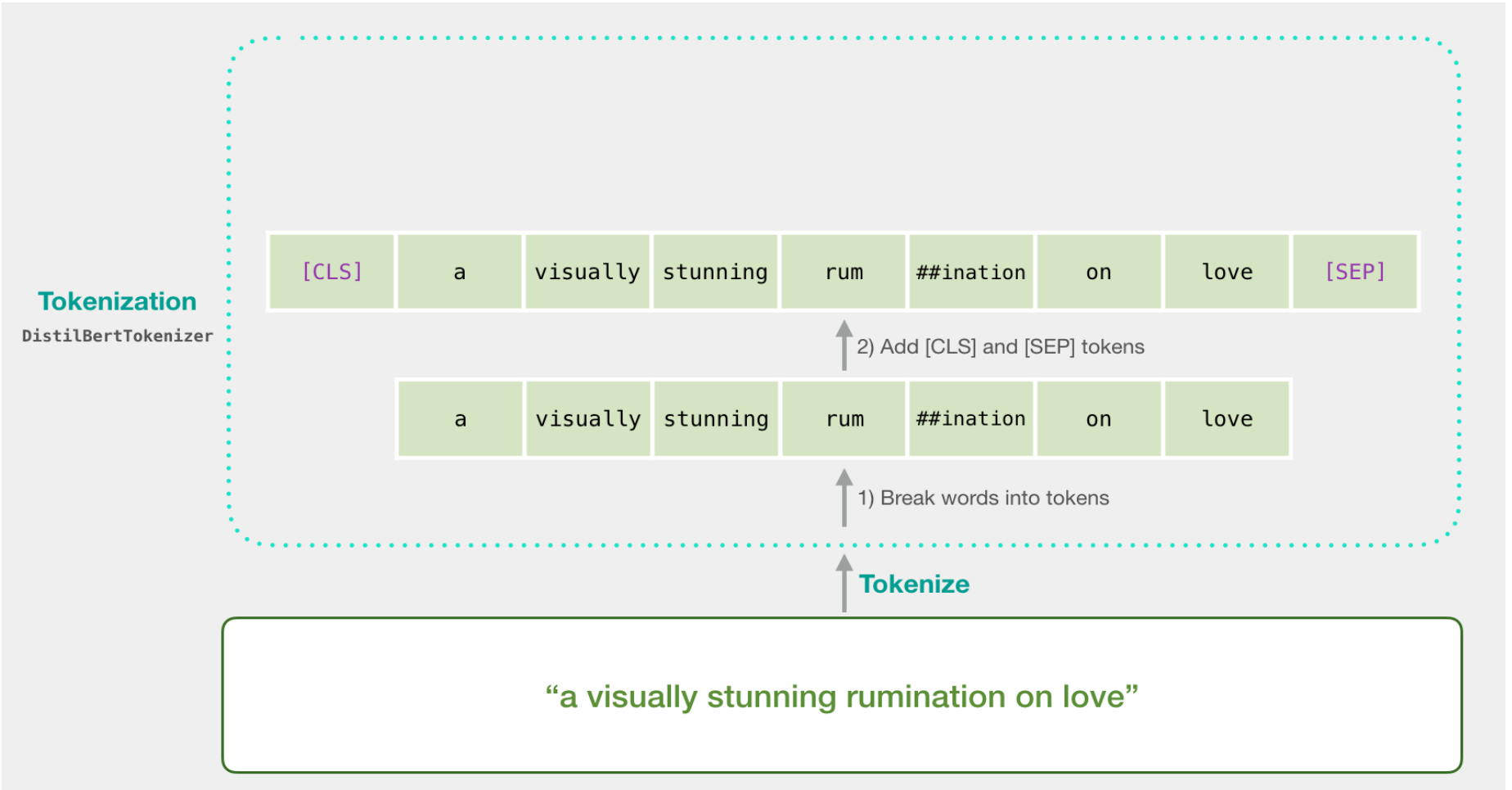
튜토리얼 개요

- 훈련셋에서 로지스틱 회귀 학습



사전학습된 모델의 예측 예시

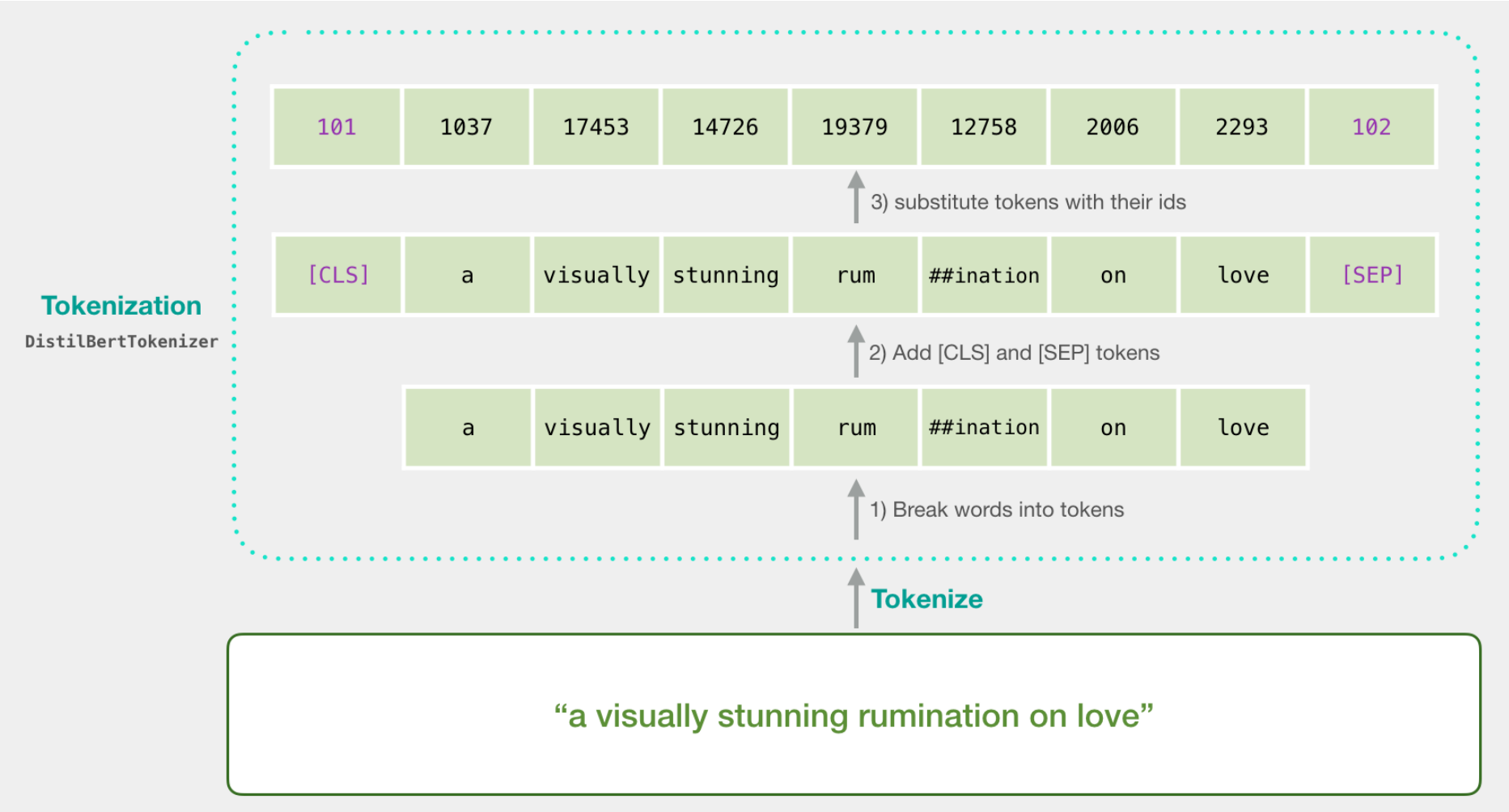
- 우선 BERT Tokenizer를 이용해 단어를 토큰으로 분리한다. 다음, 문장분류를 위한 특수 토큰을 추가한다. (처음 위치의 [CLS]와 문장 끝의 [SEP].)



사전학습된 모델의 예측 예시

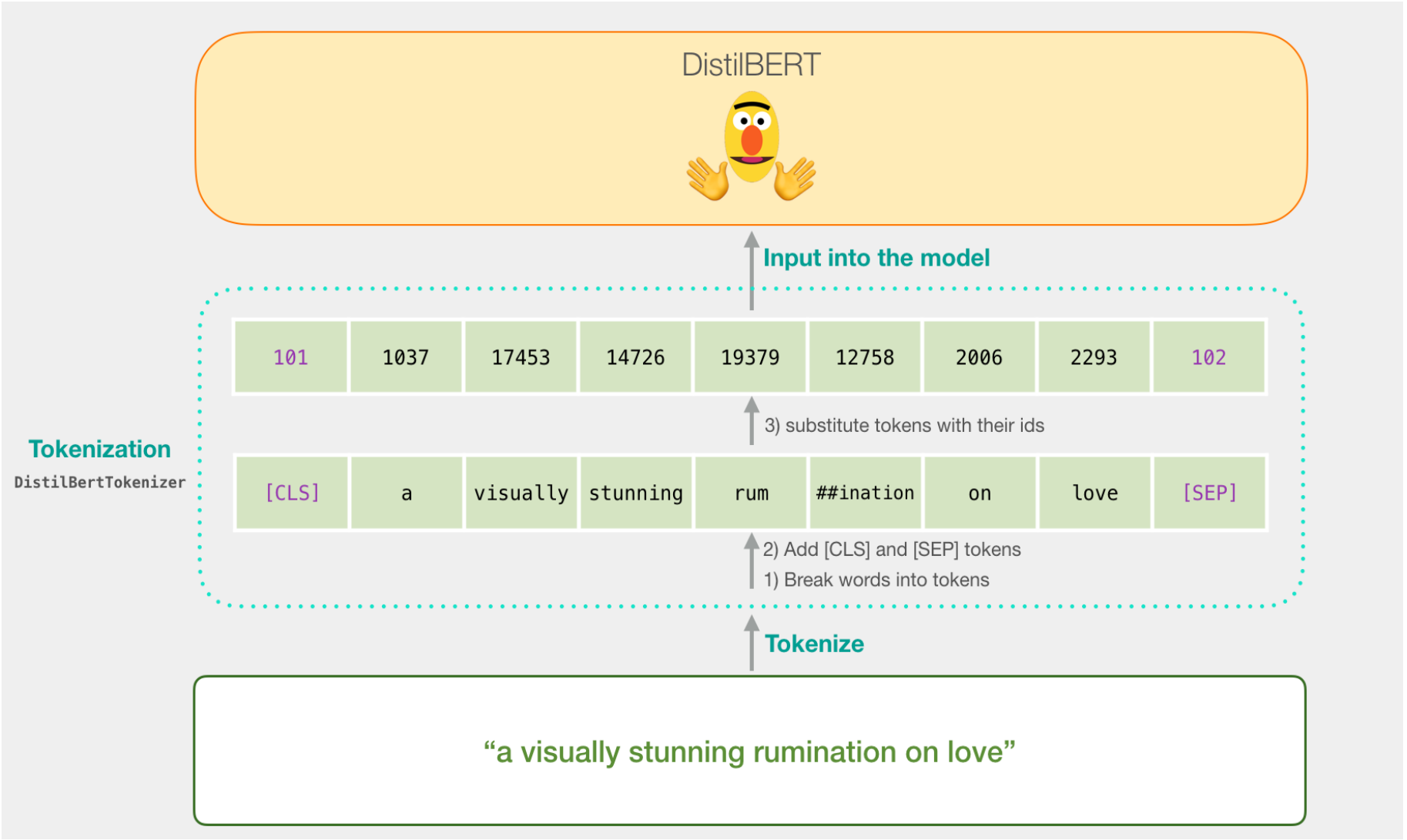
- 3번째 스텝에서 Tokenize는 각 토큰을 임베딩 테이블로부터의 id로 대체한다.

```
tokenizer.encode("a visually stunning rumination on love", add_special_tokens=True)
```



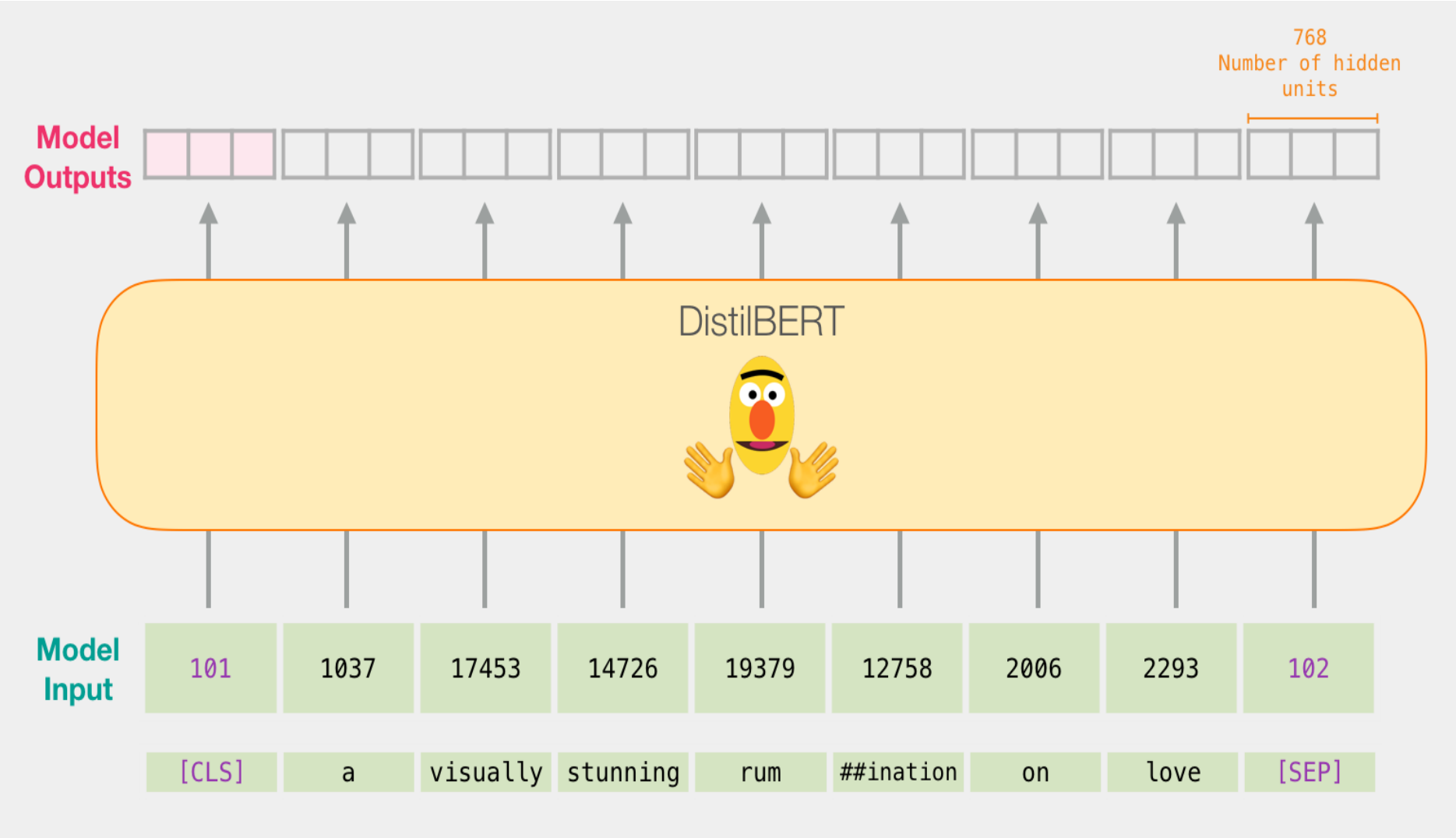
사전학습된 모델의 예측 예시

- 입력 문장은 DistilBERT에 전달될 수 있는 적절한 형태를 취한다.



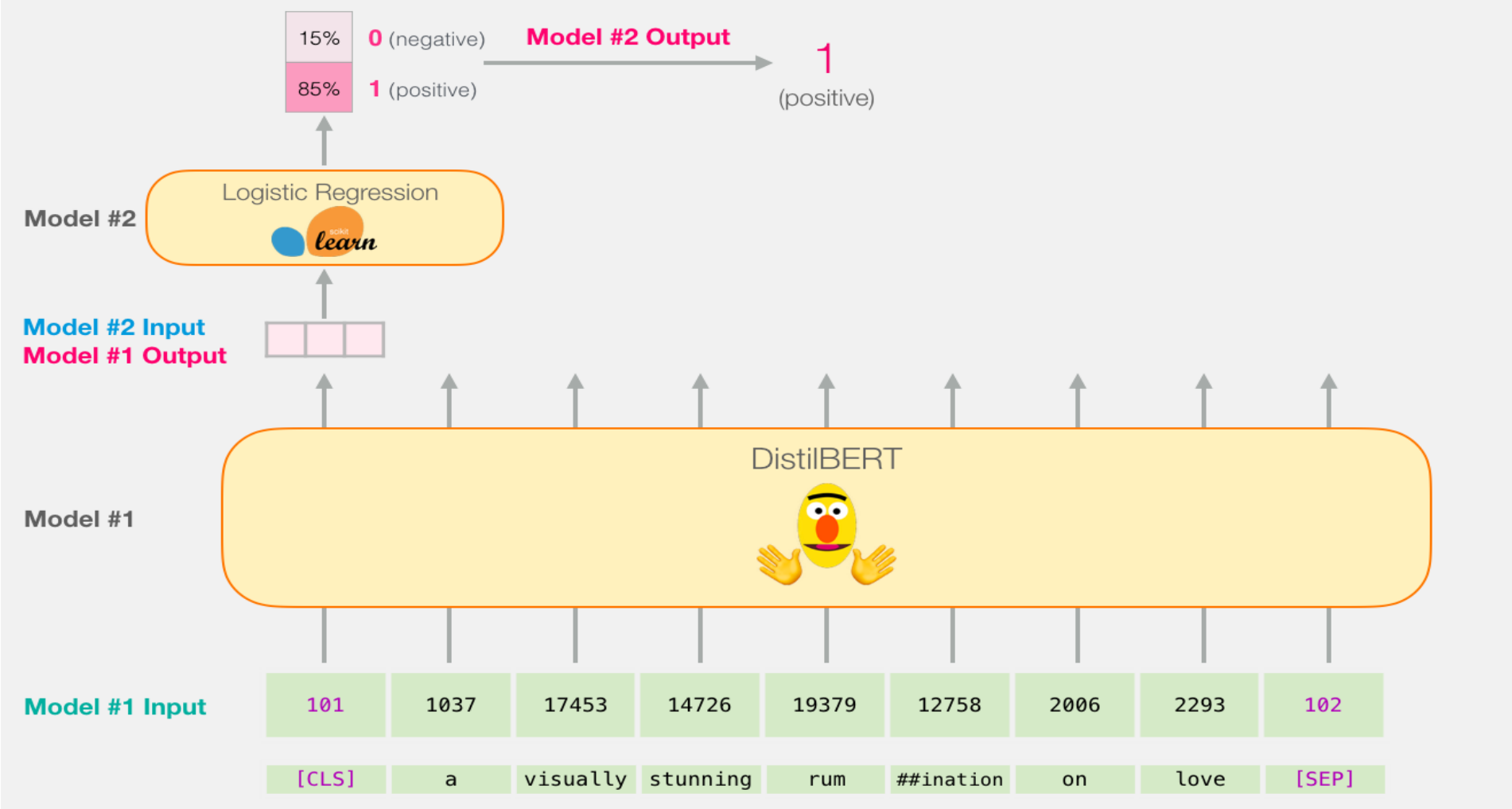
DistilBERT에서 최종 출력까지

- DistilBERT의 출력은 각 입력 토큰에 대한 벡터이며, 각 벡터는 768 개의 실수(float)로 구성된다.



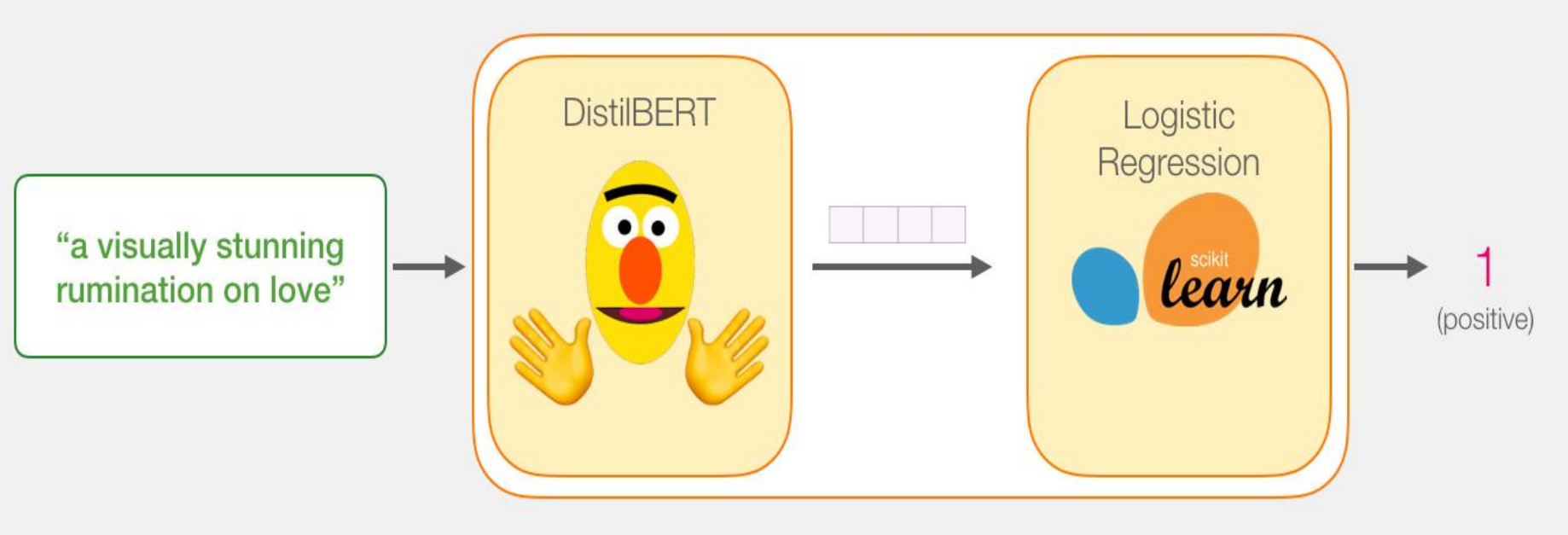
DistilBERT에서 최종 출력까지

- 문장분류이므로 첫번째 ([CLS]에 연관된)벡터만을 고려하며, 이 한 벡터가 로지스틱 회귀 모델의 입력으로 전달된다.



DistilBERT에서 최종 출력까지

- 훈련 동안 학습한 것을 기반으로 이 벡터를 분류한다. 예측 계산은 다음과 같이 될 것이다.



코드

- colab 버전 colab과 노트북버전 github을 참조하라.