

# 감성분석을 활용한 VOC System

BERT 모델 활용 도전기

발표자 : 장선영

프로젝트 기간 : 2022.11.30~2022.12.05

# 기.승.전.결

## 4일간의 고군분투

- 기 : 당초 계획 및 의의
- 승 : 구현을 위한 기술스택 조사, 실행
- 전 : 리소스 등의 현실적 어려움
- 결 : 개인으로서의 어려움, 추후 극복  
방안



기 : Motivation

# Motivation?

기업의 최우선 가치, "고객"

- 이번 프로젝트는 사업개발팀 인턴 시절의 답답함을 회고하며 시작
- 클라이언트의 '니즈' 파악이 부서 내에서 상당히 편편적인 정보를 기반, 비객관적으로 파악되는 것의 답답함
- 진짜 그들의 '니즈'는 소셜미디어 상에 여과없이 드러나 있음
- 이들의 정성적 평가를 정량화해보자! 오피니언 마이닝

# Why game data?

소셜미디어 데이터 마이닝

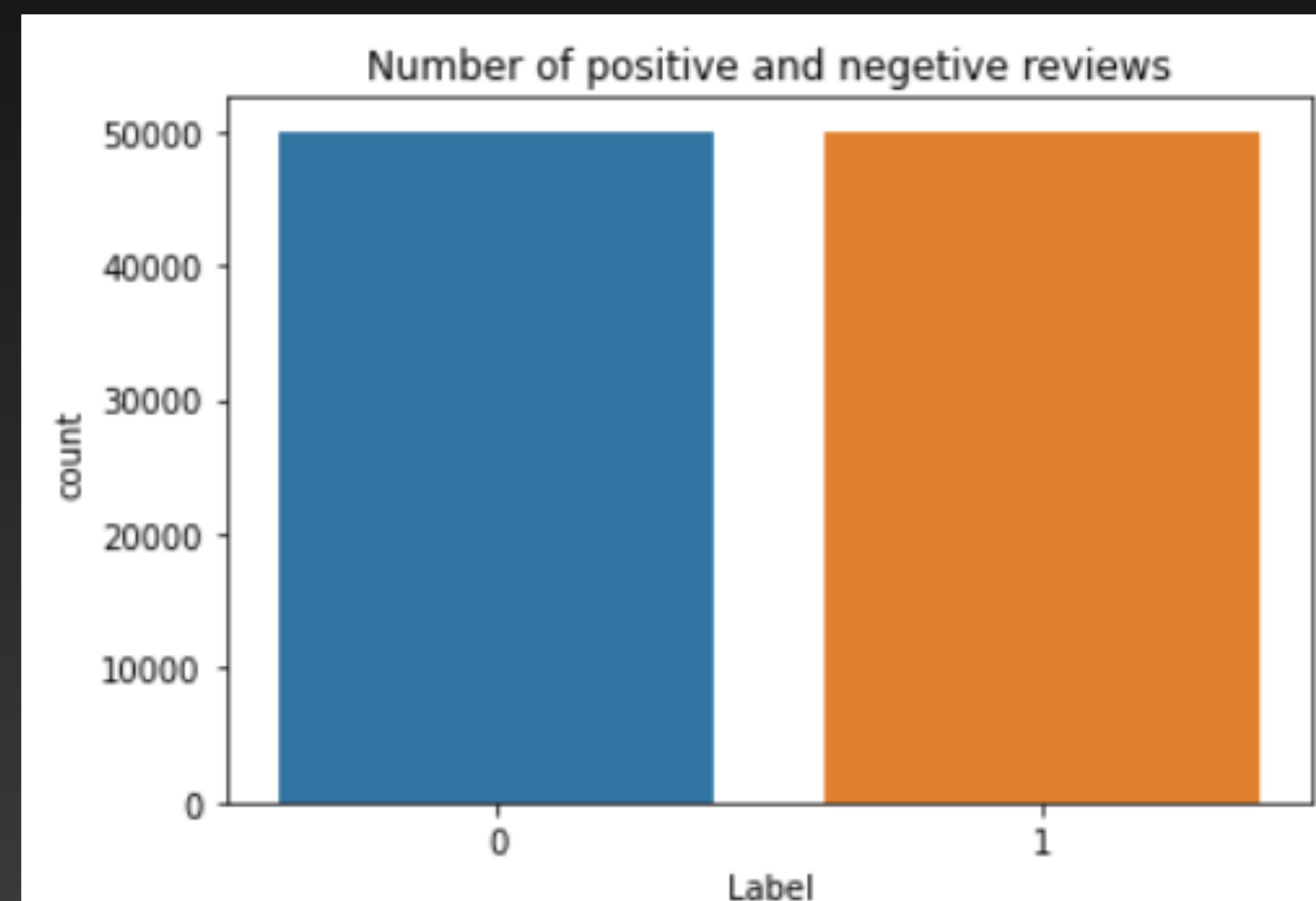
- 게임의 경우 상품 재화에 비해 소셜미디어 상에서 더 다양한 다양한 유저들의 반응을 수집 할 수 있을 것으로 기대
- 어느 도매인이든 소셜미디어 모니터링은 앞으로 더욱 중요해질 것(브랜드 매니지먼트)

# 승 : 계획 및 실행

# How?

데이터 : 스팀 한국어 리뷰 데이터  
(15만개, Labeled : 1, 0)  
모델 : LSTM, Bi-directional  
LSTM, Small BERT(en),  
BERT(base-multilingual-cased)  
*Fine-tuning* 방식

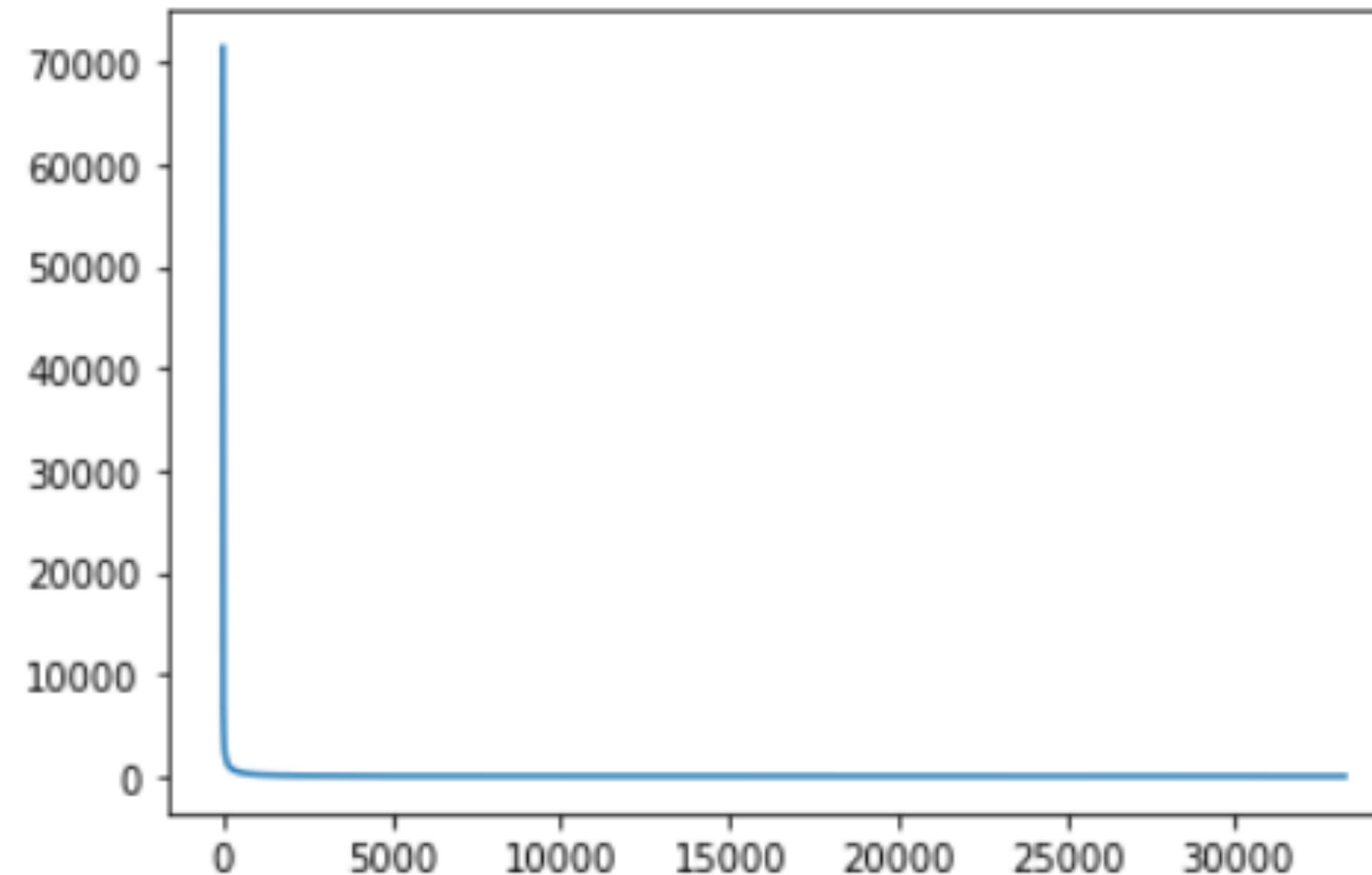
Label	reviews
0	0 노래가 너무 적음
1	0 돌겠네 진짜. 황숙아, 어크 공장 그만 돌려라. 죽는다.
2	1 막노동 체험판 막노동 하는사람인데 장비를 내가 사야돼 뭐지
3	1 차악!차악!!차악!!! 정말 아래서 왕국을 되찾을 수 있는거야??
4	1 시간 때우기에 좋음.. 도전과제는 50시간이면 다 깰 수 있어요



# Preprocessing

## Word embedding

```
# 전체 단어 빈도 수  
sorted_cnt = sorted(word_cnt, key = word_cnt.get, reverse = True)  
  
w = [word_cnt[key] for key in sorted_cnt]  
plt.plot(w)  
plt.show() # 대부분 단어가 매우 적게 사용된다.
```



```
# 가장 흔한 단어  
word_cnt = Counter(tokens)  
  
word_cnt.most_common(20)  
  
[('하다', 71474),  
 ('게임', 45198),  
 ('있다', 16394),  
 ('없다', 14719),  
 ('되다', 11321),  
 ('좋다', 10539),  
 ('같다', 10094),  
 ('재밌다', 9539),  
 ('겜', 8157),  
 ('이다', 7672),  
 ('것', 7394),  
 ('보다', 7318),  
 ('않다', 7232),  
 ('이', 7045),  
 ('아니다', 6872),  
 ('플레이', 6203),  
 ('좀', 5389),  
 ('안되다', 5134),  
 ('때', 5104),  
 ('사다', 5071)]
```

# Baseline?

## LSTM

- 감성분석 공부로 예제를 따라 만든 모델
- 별 큰 기대없이 사용
- Tokenizer : KoNLP - twitter
- Test dataset 성능 : .076

Model : "model"

Layer (type)	Output Shape	Param #
inputs (InputLayer)	[ (None, 1500) ]	0
embedding (Embedding)	(None, 1500, 50)	500000
lstm (LSTM)	(None, 64)	29440
FC1 (Dense)	(None, 256)	16640
activation (Activation)	(None, 256)	0
dropout (Dropout)	(None, 256)	0
out_layer (Dense)	(None, 1)	257
activation_1 (Activation)	(None, 1)	0

Total params: 546,337

Trainable params: 546,337

Non-trainable params: 0

# 전 : 계획과 현실의 괴리

# 위기 1

## 소잡는 칼 닦 잡는데 사용?

- 8시간 겨우 시간들여 완성한 Bert 모델 성능이 Baseline을 하회(.72%)
- 코랩 GPU 할당량 초과로 더는 무거운 모델을 학습하는 것이 불가
- BERT model 2종류에 대한 학습 중 성능 비교
- Small BERT en : val accuracy가 .50에서 증가하지 않아 학습 중단
- BERT(base-multilingual-cased) : 마찬가지로 .73에서 증가 x

결국 CPU로 Base model tuning...  
(4시간 소요...)

77%

최종 Testset Accuracy  
모델 : Bi-directional LSTM

# 결 : 한계점 및 의의

그래도 이건 배웠다...

## 의의

Bert 모델의 한계와 추후 활용 방안에 대한 인사이트

- 데이터가 적을 땐 전이학습, 데이터가 많고 GPU 넉넉하면 fine-tuning
- BERT 모델 구조에 대해 이해하고 활용해봄
- 딥러닝 하이퍼 파라미터 튜닝 및 NLP 공부해보는 계기
- KoNLP 토크나이저 활용

## 추후 보완할 점

- Topic modeling을 활용하여 각 리뷰에 대한 토픽을 분류,
- KoBERT 전이학습을 이용하여 각 토픽에 대한 감성분석
- → 단순 감성 분석에서 그치는 것이 아닌 주제 별 감성을 분류함으로써, 더욱 구체적인 고객의 니즈 파악 가능, 리소스 한계를 인지하고 전이학습을 통한 모델 최적화
- → 실제 VOC 서비스를 위한 GUI 구현





# 감사합니다!

피드백은 발표자의 소중한 자산이 됩니다  
간단하게나마 부탁드립니다! 😊