# 감성 분석을 통한 호감도 예측

2022.04.06

김수영

MLCL Kyungpook National University

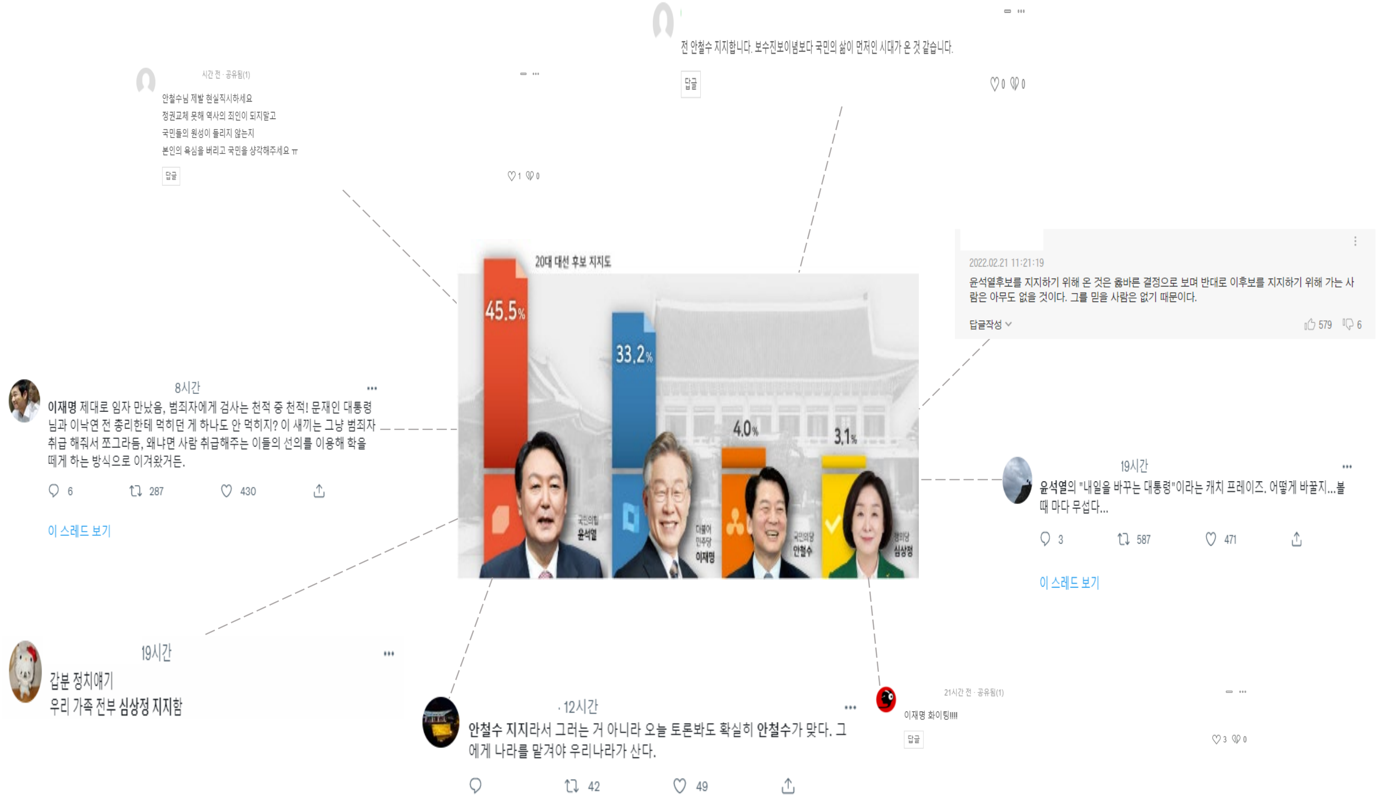
# CONTENTS

1. Intro
2. Data
3. Experiment
4. Analysis
5. Conclusion
6. Future works

01

# Intro

# Project Name: 감성적인 투표



시간 전 · 공유됨(1)

안철수님 제발 현실직시하세요
정권교체 못해 역사의 죄인이 되지말고
국민들의 원성이 들리지 않는지
본인의 욕심을 버리고 국민을 생각해주세요 ㅠ

답글 ♡ 1 0

전 안철수 지지합니다. 보수진보이념보다 국민의 삶이 먼저인 시대가 온 것 같습니다.

답글 ♡ 0 0

2022.02.21 11:21:19

윤석열후보를 지지하기 위해 온 것은 옳바른 결정으로 보며 반대로 이후보를 지지하기 위해 가는 사람은 아무도 없을 것이다. 그를 믿을 사람은 없기 때문이다.

답글작성 579 6

8시간

**이재명** 제대로 임자 만났음, 범죄자에게 검사는 천적 중 천적! 문재인 대통령님과 이낙연 전 총리한테 먹히던 게 하나도 안 먹히지? 이 새끼는 그냥 범죄자 취급 해줘서 쪼그라듬, 왜냐면 사람 취급해주는 이들의 선의를 이용해 학을 떼게 하는 방식으로 이겨왔거든.

6 287 430

이 스레드 보기

19시간

**윤석열**의 "내일을 바꾸는 대통령"이라는 캐치 프레이즈. 어떻게 바꿀지...볼 때 마다 무섭다...

3 587 471

이 스레드 보기

19시간

갑분 정치얘기
우리 가족 전부 심상정 지지함

· 12시간

**안철수 지지**라서 그러는 거 아니라 오늘 토론봐도 확실히 **안철수**가 맞다. 그에게 나라를 맡겨야 우리나라가 산다.

42 49

21시간 전 · 공유됨(1)

이재명 화이팅!!!!

답글 ♡ 3 0

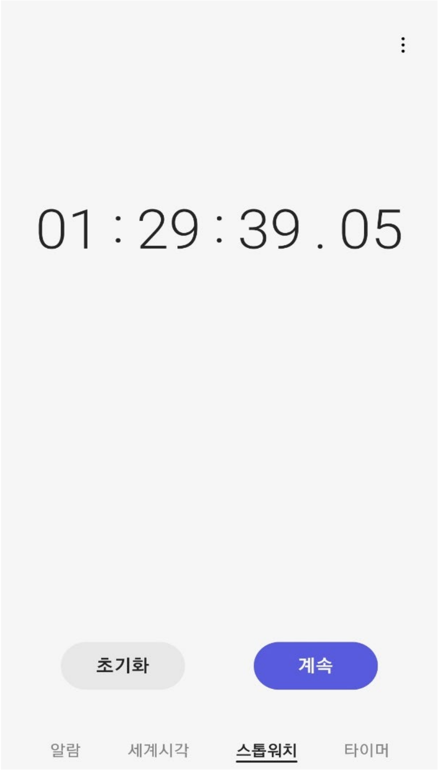
# Intro

1K

MLCL Kyungpook National University

# Intro

1K



01 : 29 : 39 . 05

# Goals

- Text–based unseen data 활용 연구

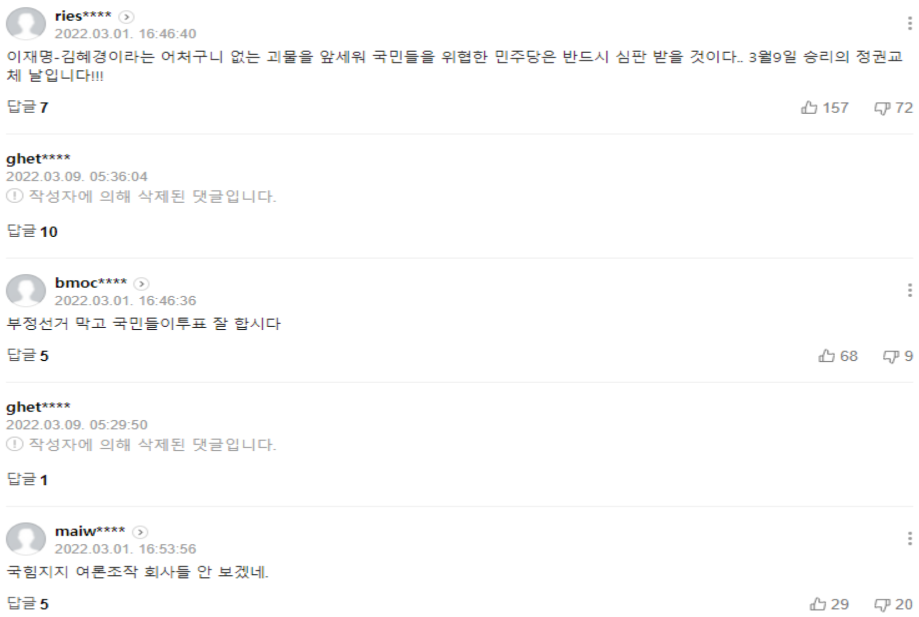
- 감성 분석을 통한 후보 호감도 예측

02

# Data

# Data Collection

NAVER



ries****
2022.03.01. 16:46:40
이재명-김혜경이라는 어처구니 없는 괴물을 앞세워 국민들을 위협한 민주당은 반드시 심판 받을 것이다.. 3월9일 승리의 정권교체 날입니다!!!
답글 7 👍 157 👎 72

ghet****
2022.03.09. 05:36:04
ⓘ 작성자에 의해 삭제된 댓글입니다.
답글 10

bmoc****
2022.03.01. 16:46:36
부정선거 막고 국민들이투표 잘 합시다
답글 5 👍 68 👎 9

ghet****
2022.03.09. 05:29:50
ⓘ 작성자에 의해 삭제된 댓글입니다.
답글 1

maiw****
2022.03.01. 16:53:56
국힘지지 여론조작 회사들 안 보겠네.
답글 5 👍 29 👎 20

**Date: 2/1 ~ 3/4**

**Data count: 2730K**

**Method: Web Crawling**

# Data Pre-processing

2730K ⇨

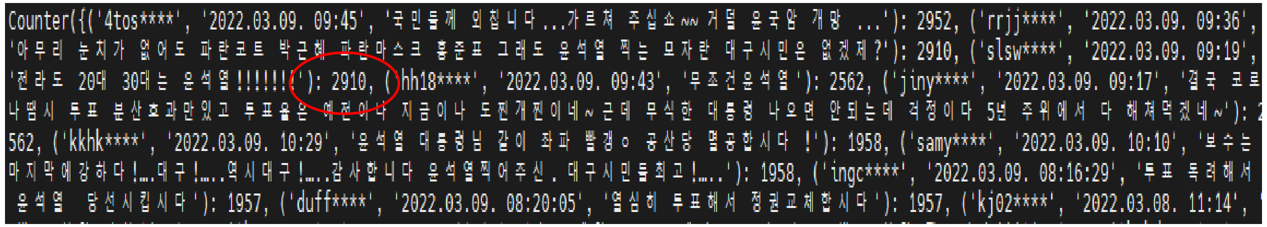
Data classification

| | |
|---|---|
| 이재명 | 336,049 |
| 윤석열 | 386,539 |
| 중복 | 104,414 |
| 총 합 | 827K |

⇨

Deduplication

| | |
|---|---|
| 이재명 | 11,675 |
| 윤석열 | 16,140 |
| 중복 | 4,248 |
| 총 합 | 32K |



```
Counter({('4tos****', '2022.03.09. 09:45', '국민들께 외칩니다...가르쳐 주십쇼ㅆㅆ 거덜 윤국양 개망 ...'): 2952, ('rrjj****', '2022.03.09. 09:36',
'아무리 눈치가 없어도 파란코트 박근혜 파란마스크 홍준표 그래도 윤석열 찍는 모자란 대구시민은 없겠제?'): 2910, ('slsw****', '2022.03.09. 09:19',
'전라도 20대 30대는 윤석열!!!!!!'): 2910, ('hh18****', '2022.03.09. 09:43', '무조건윤석열'): 2562, ('jiny****', '2022.03.09. 09:17', '결국 코로
나뗌시 투표 분산효과만있고 투표율은 예전이나 지금이나 도찐개찐이네~근데 무식한 대통령 나오면 안되는데 걱정이다 5년 주위에서 다 해쳐먹겠네~'): 2
562, ('kkhk****', '2022.03.09. 10:29', '윤석열 대통령님 같이 좌파 빨갱ㅇ 공산당 멸공합시다 !'): 1958, ('samy****', '2022.03.09. 10:10', '보수는
마지막에강하다!….대구!…..역시대구!…..감사합니다 윤석열찍어주신. 대구시민들최고!…..'): 1958, ('ingc****', '2022.03.09. 08:16:29', '투표 독려해서
 윤석열  당선시킵시다'): 1957, ('duff****', '2022.03.09. 08:20:05', '열심히 투표해서 정권교체합시다'): 1957, ('kj02****', '2022.03.08. 11:14', '
```

MLCL Kyungpook National University

# Data

▷ Naver news comment data

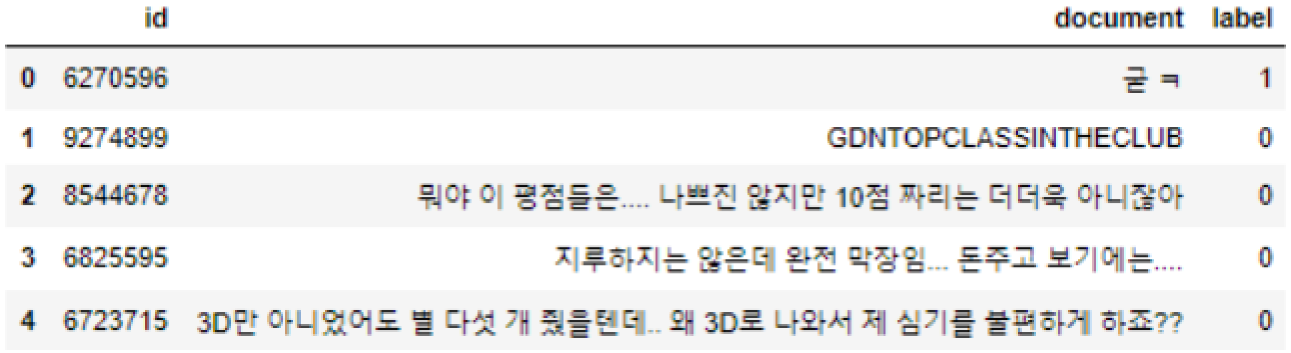
| | |
|---|---|
| 이재명 | 11,675 |
| 윤석열 | 16,140 |
| 중복 | 4,248 |
| 총 합 | 32,063 |

⇨

| | 윤석열 | 이재명 | 총합 |
|---|---|---|---|
| Train | 9,000 | 9,000 | 18,000 |
| Test | 1,000 | 1,000 | 2,000 |

▷ Naver movie review data

| | |
|---|---|
| Train | 150K |
| Test | 50K |

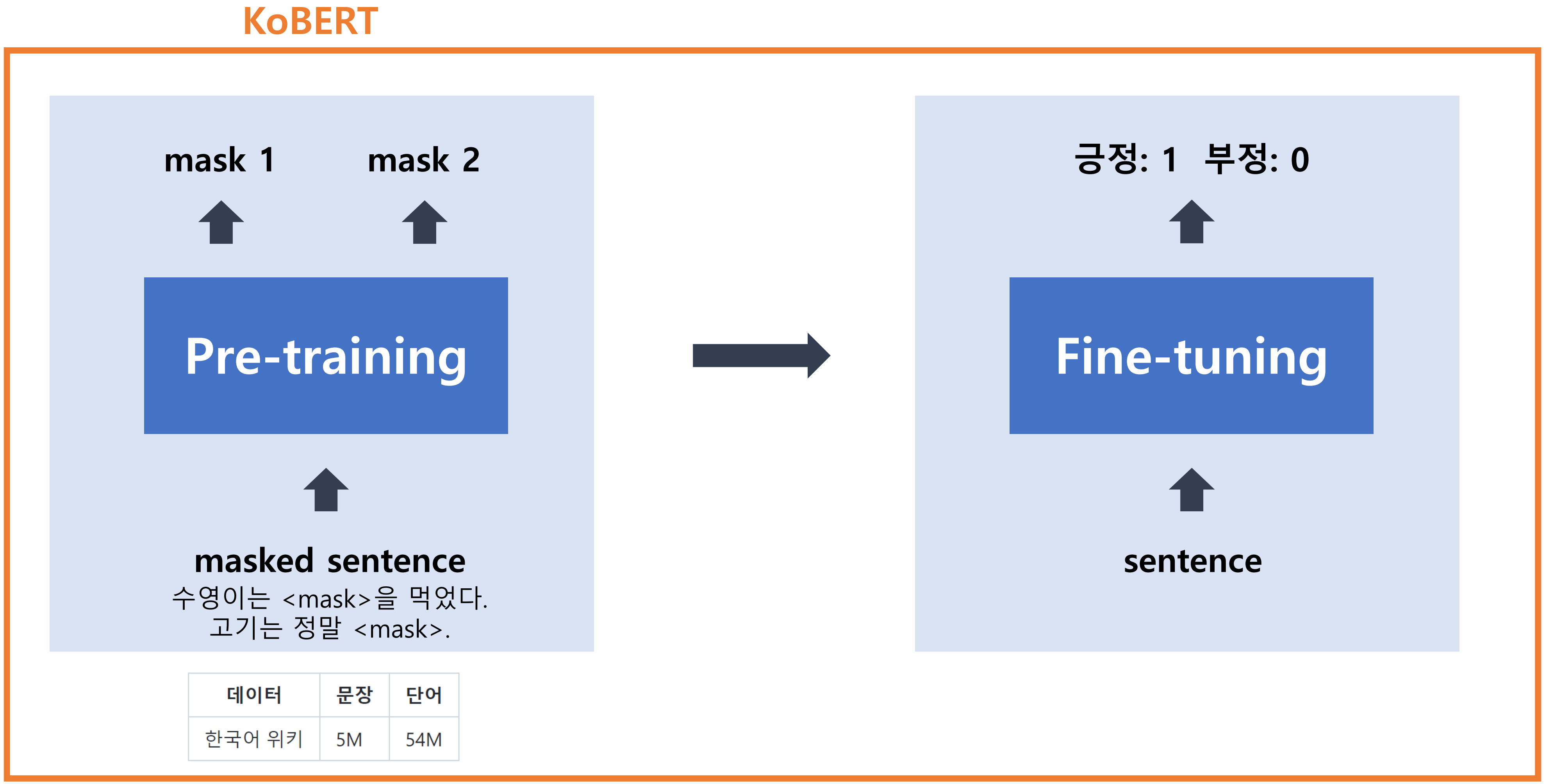


| | id | document | label |
|---|---|---|---|
| 0 | 6270596 | 굳 ㅋ | 1 |
| 1 | 9274899 | GDNTOPCLASSINTHECLUB | 0 |
| 2 | 8544678 | 뭐야 이 평점들은.... 나쁘진 않지만 10점 짜리는 더더욱 아니잖아 | 0 |
| 3 | 6825595 | 지루하지는 않은데 완전 막장임... 돈주고 보기에는.... | 0 |
| 4 | 6723715 | 3D만 아니었어도 별 다섯 개 줬을텐데.. 왜 3D로 나와서 제 심기를 불편하게 하죠?? | 0 |

03

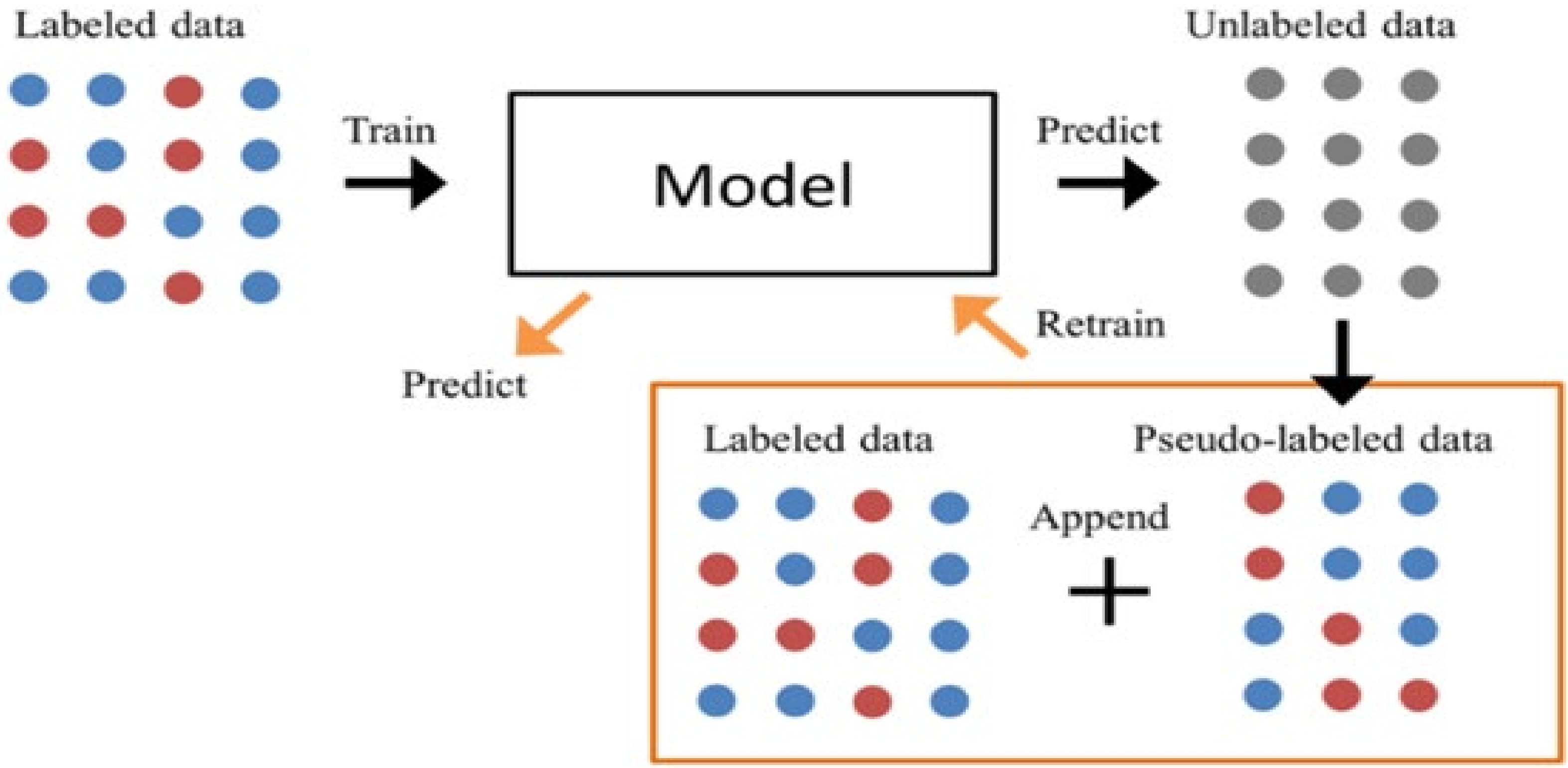
# Experiment

# Method1. Pre-trained model (KoBERT)

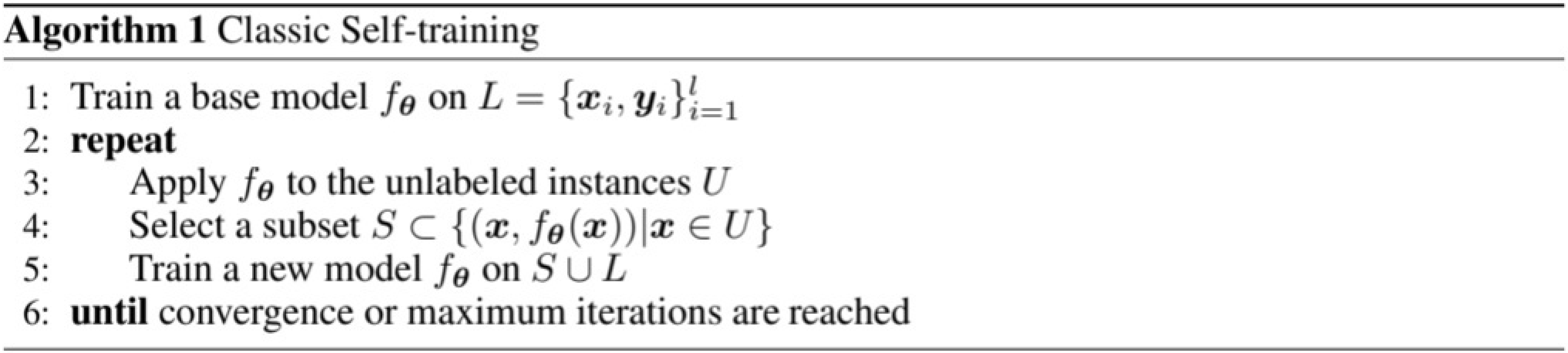


| 데이터 | 문장 | 단어 |
|---|---|---|
| 한국어 위키 | 5M | 54M |

# Method2. Pre-trained model + Self-training



# Method2. Pre-trained model + Classic Self-training



**Algorithm 1** Classic Self-training

```
Train a base model $f_\theta$ on $L = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{l}$
repeat
    Apply $f_\theta$ to the unlabeled instances $U$
    Select a subset $S \subset \{(\boldsymbol{x}, f_\theta(\boldsymbol{x})) | \boldsymbol{x} \in U\}$
    Train a new model $f_\theta$ on $S \cup L$
until convergence or maximum iterations are reached
```

From [1]

[1] He, Junxian, et al. "Revisiting self-training for neural sequence generation." arXiv preprint arXiv:1909.13788 (2019).

# Data

| | Pre-trained model (KoBERT) | Self-training |
|---|---|---|
| Labeled data<br>(movie review) | 200K | 200K |
| Unlabeled data<br>(news comment) | 18K | 18K |
| Test<br>(news comment) | 2K | 2K |

# Parameter

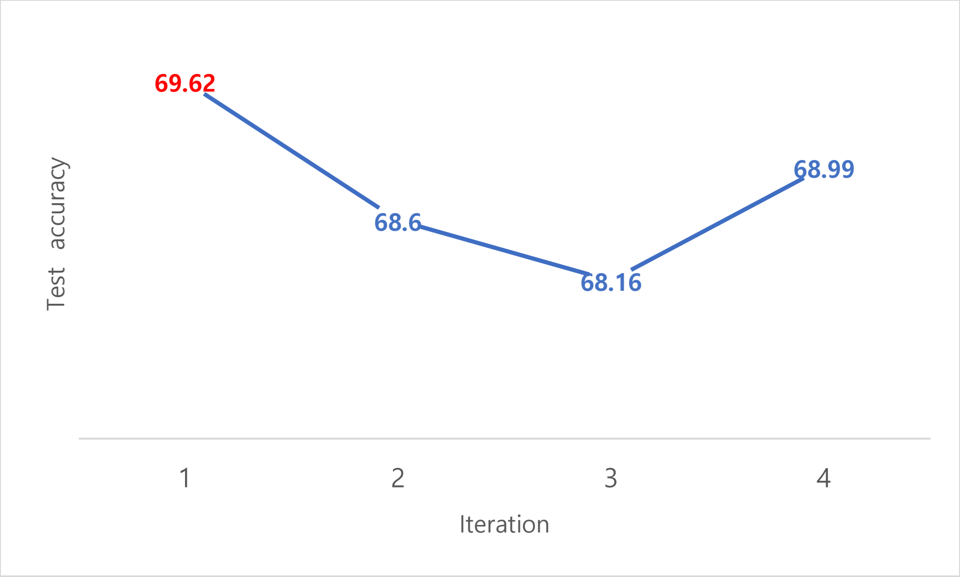
| | Pre-trained model (KoBERT) | Self-training |
|---|---|---|
| Batch size | 64 | 64 |
| Epoch | 5 | 5 |
| Learning rate | 1e-4 | 1e-4 |
| Dropout | 0.5 | 0.5 |
| Iteration | - | 4 |
| Optimizer | AdamW | AdamW |
| Criterion | CrossEntropyLoss | CrossEntropyLoss |

# Experiment Result

## Method1. Pre-trained model (KoBERT)

68.16 %

## Method2. Classic Self-training



04

# Analysis

# Analysis

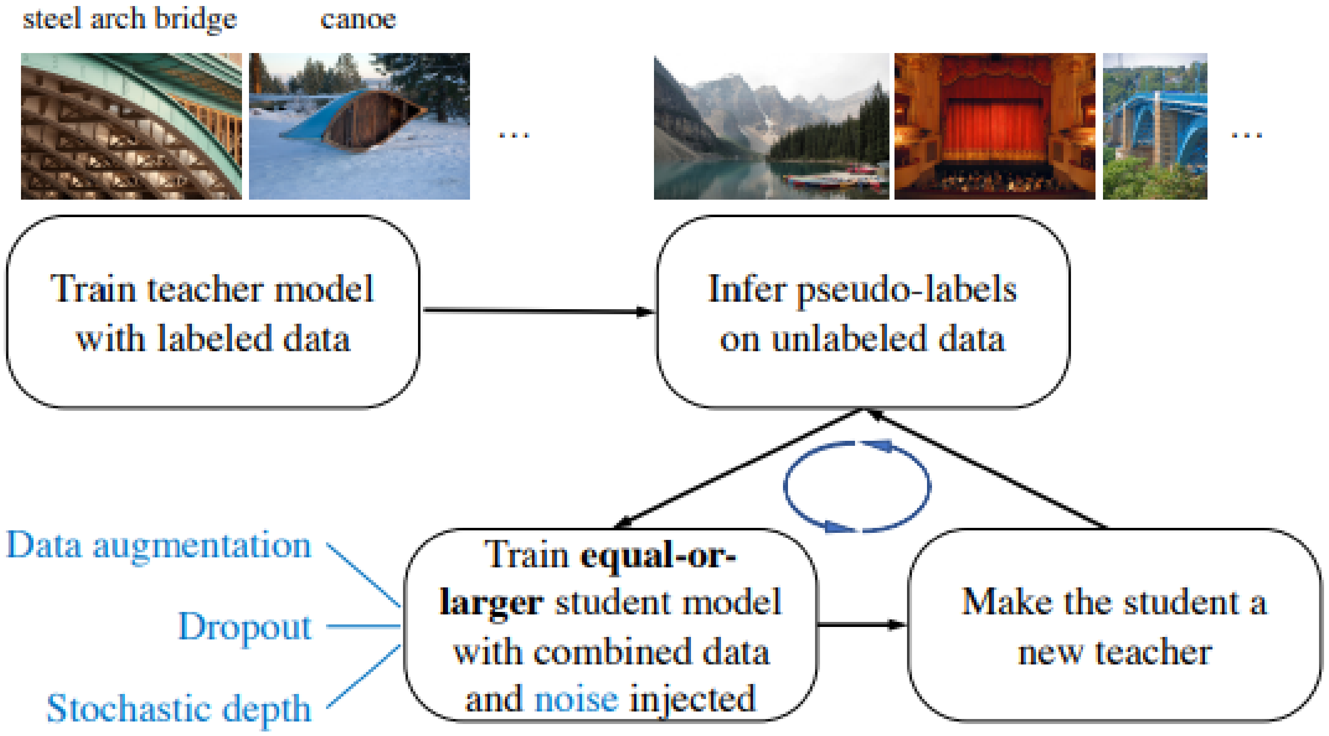
1. Self-training with Noisy Student

2. Data ratio

3. Labeled data change

4. Test data change and F1-score measure

# 1. Self–training with Noisy Student



steel arch bridge

canoe

...

...

Train teacher model with labeled data

Infer pseudo-labels on unlabeled data

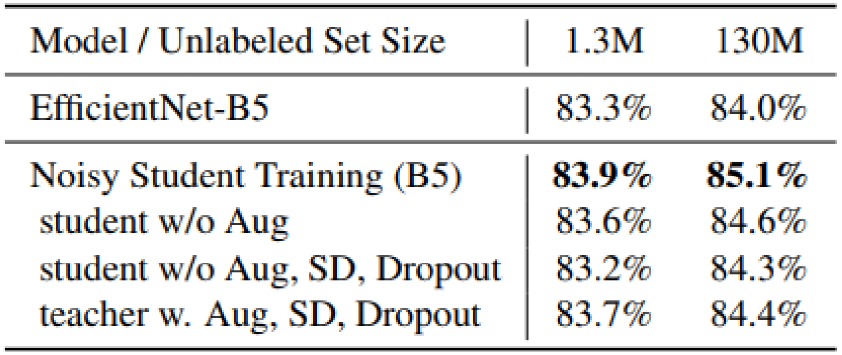
Train **equal-or-larger** student model with combined data and noise injected

Make the student a new teacher

Data augmentation

Dropout

Stochastic depth

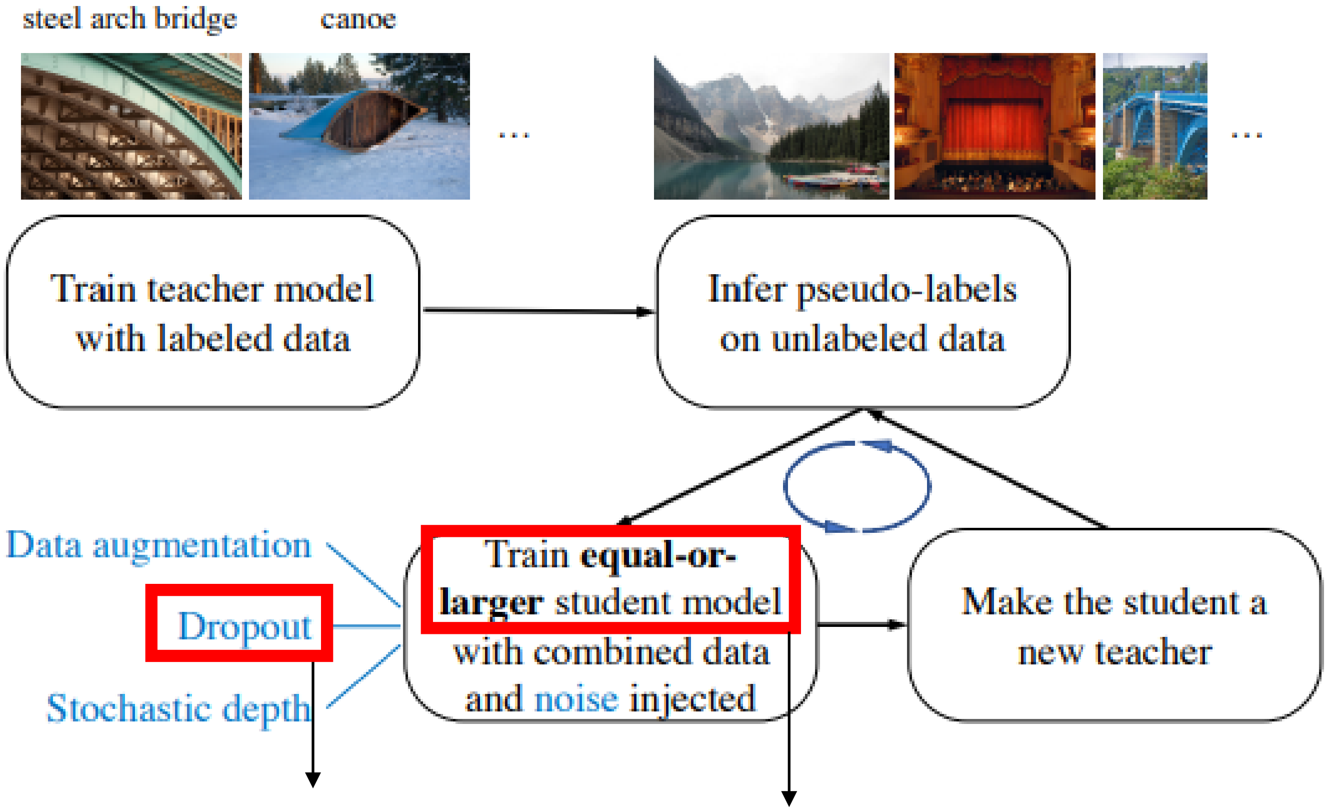


| Model / Unlabeled Set Size | 1.3M | 130M |
|---|---|---|
| EfficientNet-B5 | 83.3% | 84.0% |
| Noisy Student Training (B5) | **83.9%** | **85.1%** |
| student w/o Aug | 83.6% | 84.6% |
| student w/o Aug, SD, Dropout | 83.2% | 84.3% |
| teacher w. Aug, SD, Dropout | 83.7% | 84.4% |

From [1]

[1] Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

# 1. Self–training with Noisy Student



Train teacher model with labeled data

Infer pseudo-labels on unlabeled data

Data augmentation

Dropout

Stochastic depth

Train **equal-or-larger** student model with combined data and noise injected

Make the student a new teacher

KoBERT
dropout = 0.2

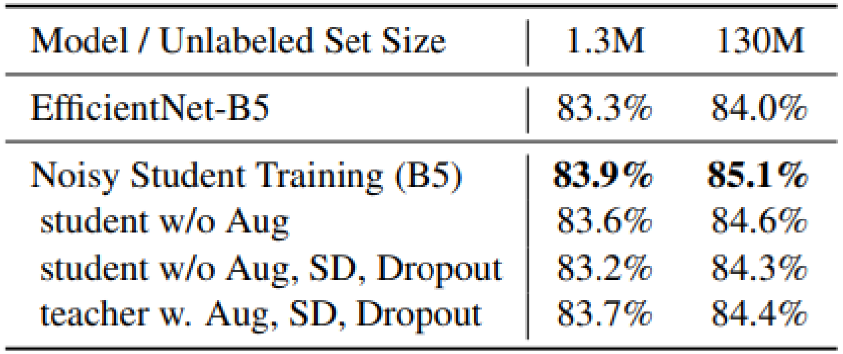
number of data

(1) 4500

(2) 9000

(3) 13500

(4) 18000



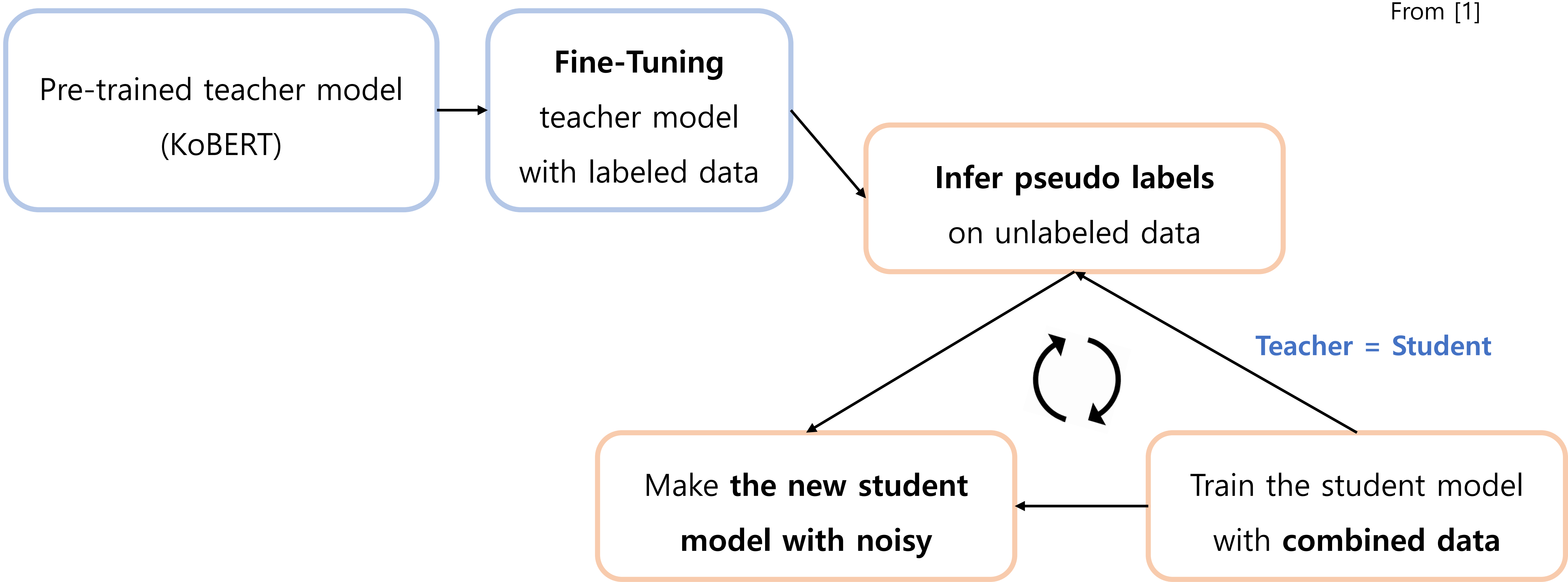
| Model / Unlabeled Set Size | 1.3M | 130M |
|---|---|---|
| EfficientNet-B5 | 83.3% | 84.0% |
| Noisy Student Training (B5) | **83.9%** | **85.1%** |
| student w/o Aug | 83.6% | 84.6% |
| student w/o Aug, SD, Dropout | 83.2% | 84.3% |
| teacher w. Aug, SD, Dropout | 83.7% | 84.4% |

From [1]

[1] Xie, Qizhe, et al. "Self-training with noisy student improves imagenet classification." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

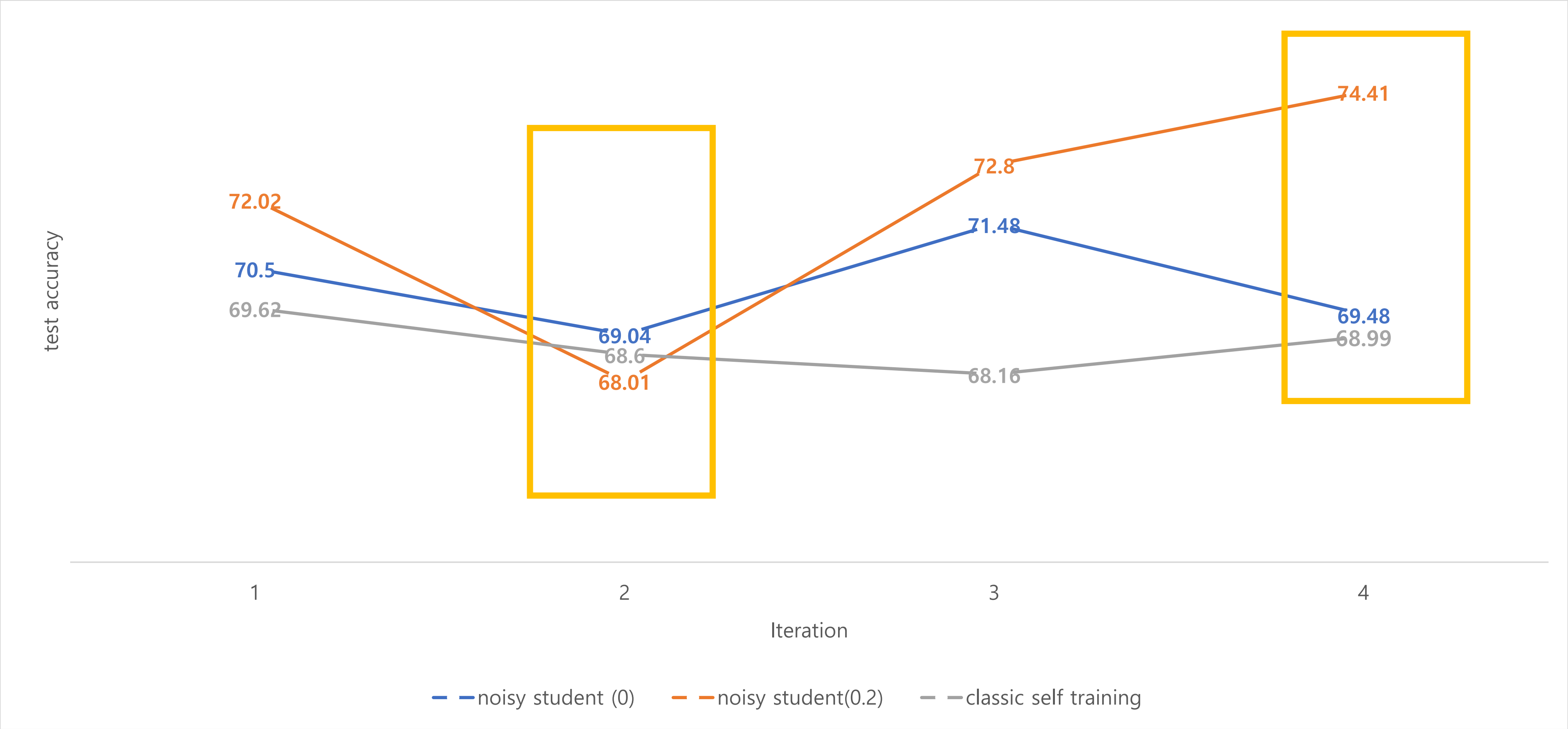
# 1. Self–training with Noisy Student

From [1]

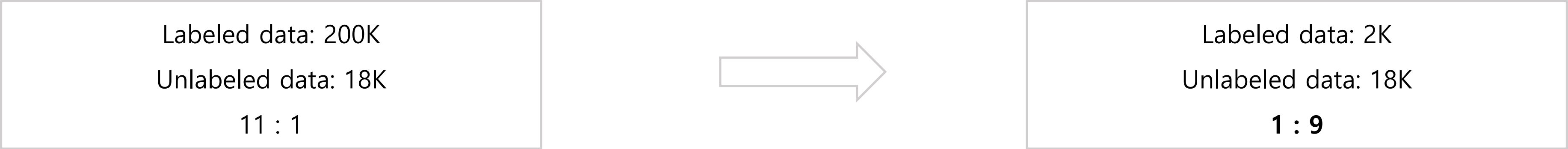


[1] Zhang, Yu, et al. "Pushing the limits of semi-supervised learning for automatic speech recognition." *arXiv preprint arXiv:2010.10504* (2020).

# 1. Self–training with Noisy Student - Result



# 2. Data ratio



Labeled data: 200K

Unlabeled data: 18K

11 : 1

Labeled data: 2K

Unlabeled data: 18K

**1 : 9**

▷ Infer pseudo labels on **accumulated** unlabeled data

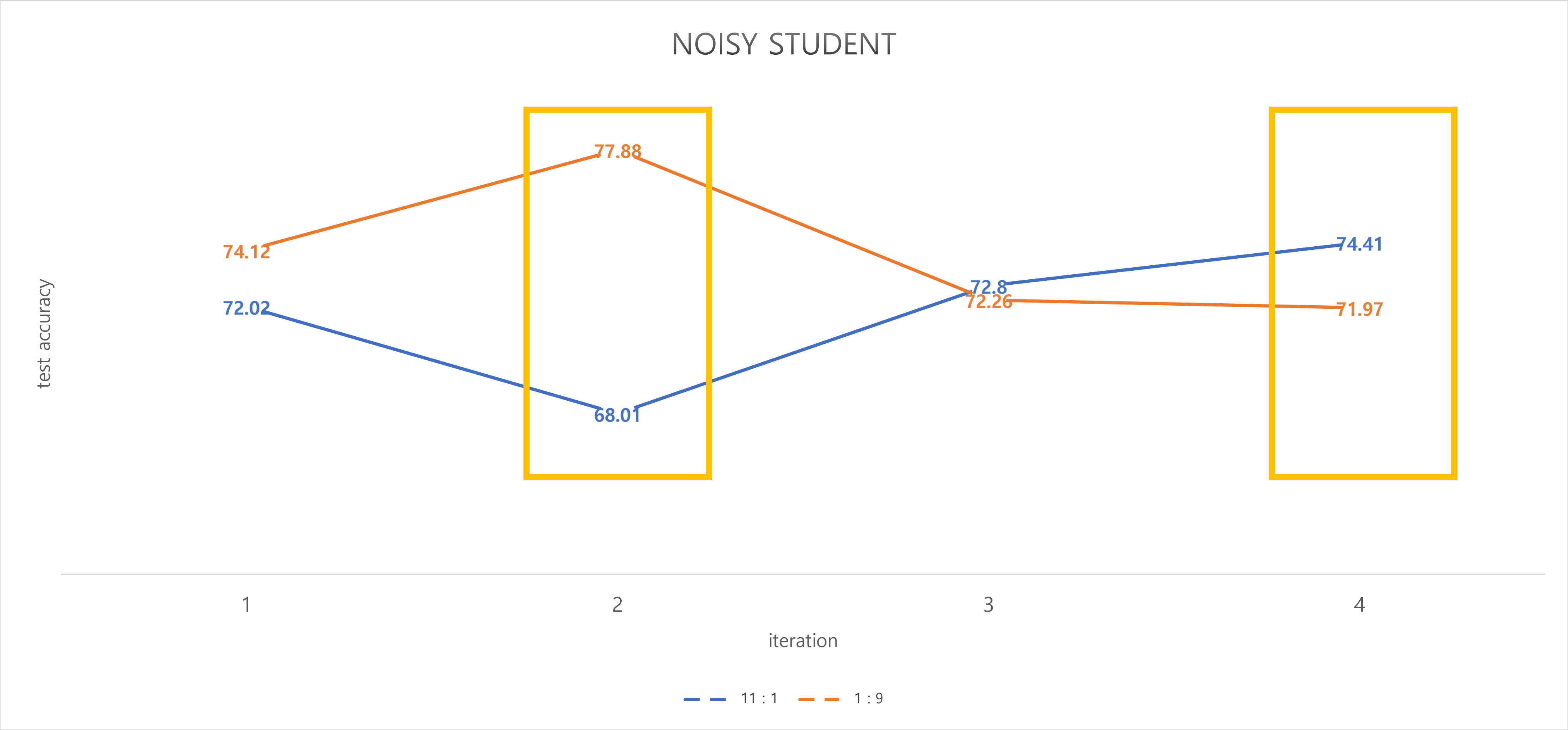
(1) 500 : 4500

(2) 1000 : 4500 + 4500

(3) 1500 : 9000 + 4500

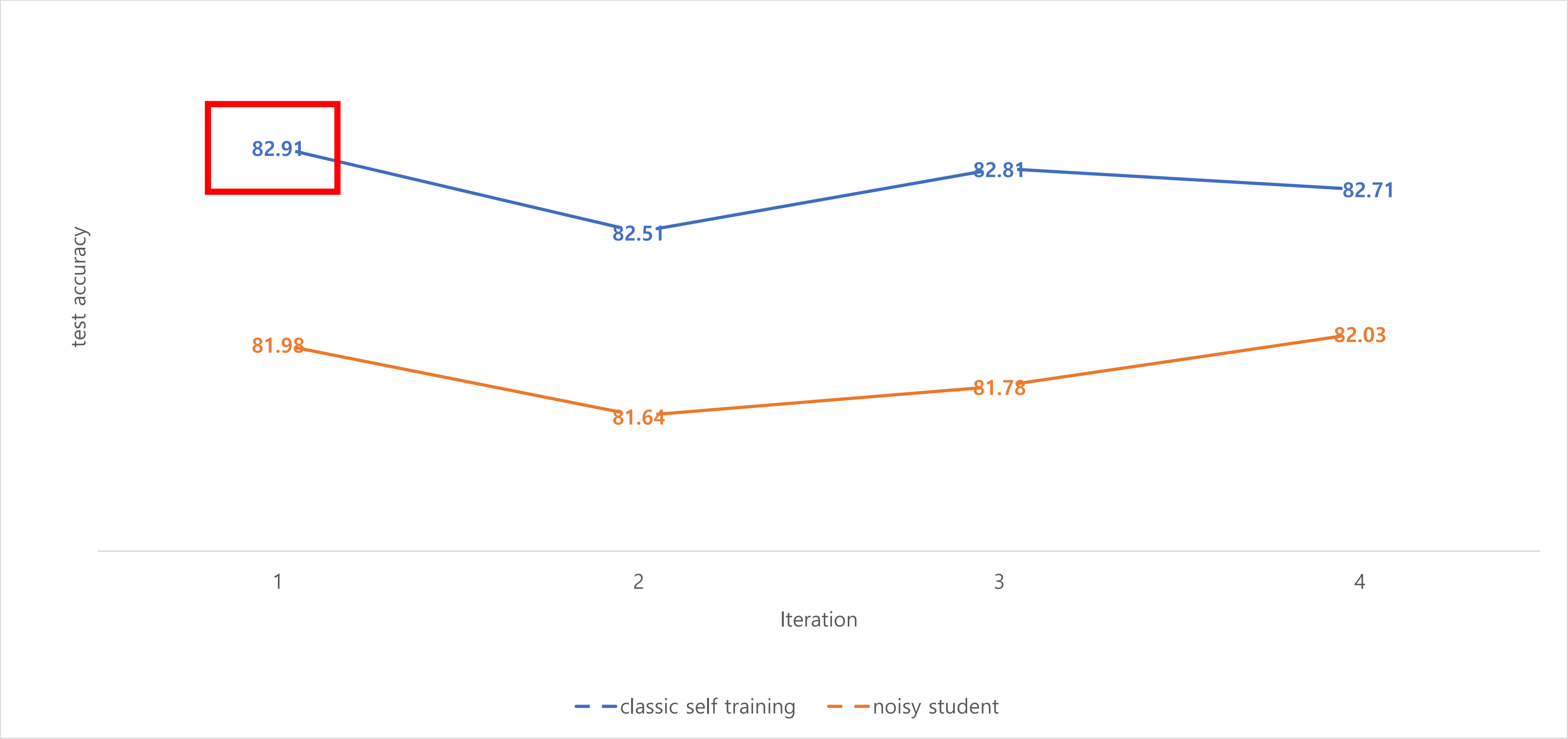
(4) 2000 : 13500 + 4500

# 2. Data ratio - Result



MLCL Kyungpook National University

# 3. Labeled Data Change (1 : 9)



Kyungpook National University

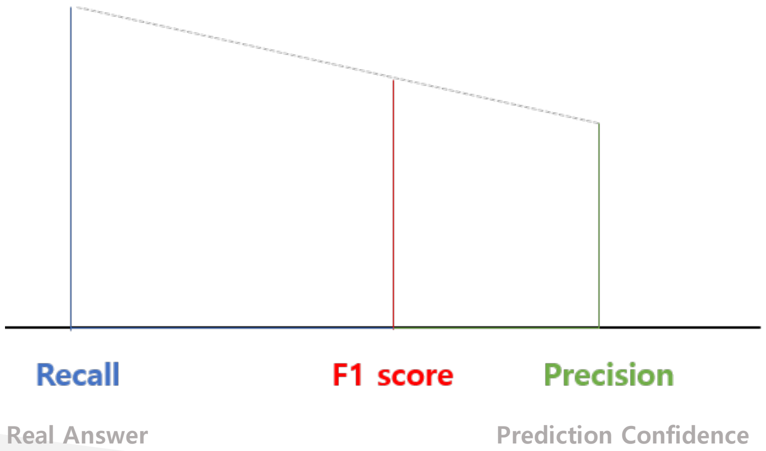
# 4. Test data change and F1-score measure

▷ Test data change

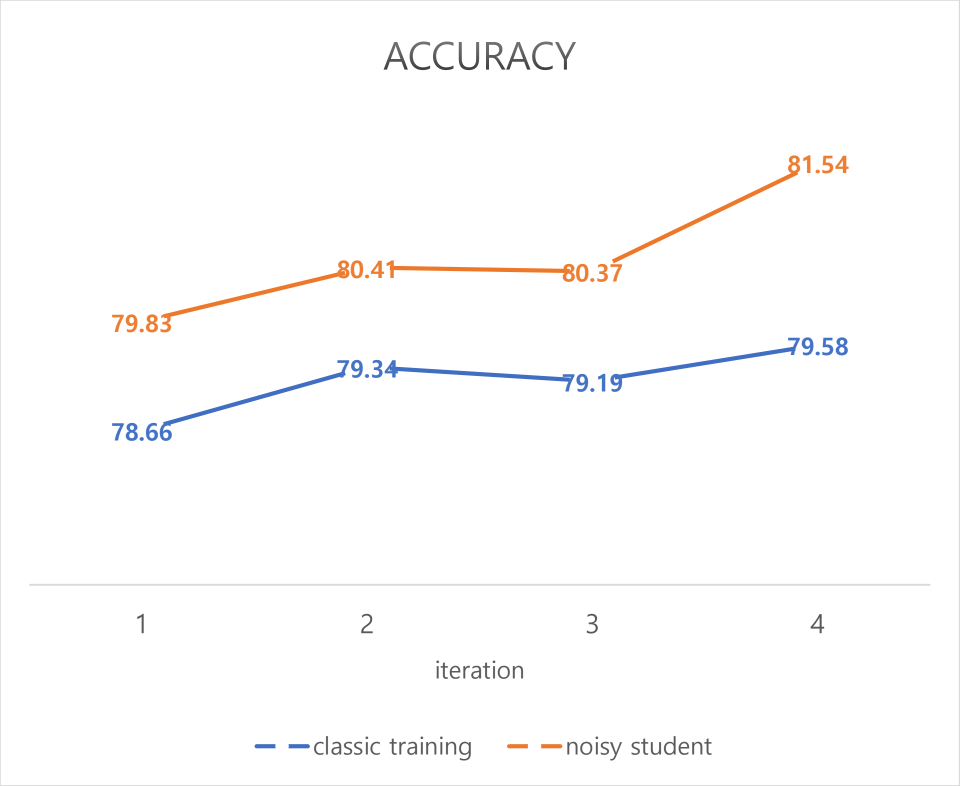
| 0 | 1 |
|---|---|
| 71.6% | 28.4% |

| 0 | 1 |
|---|---|
| 58.05% | 41.95% |

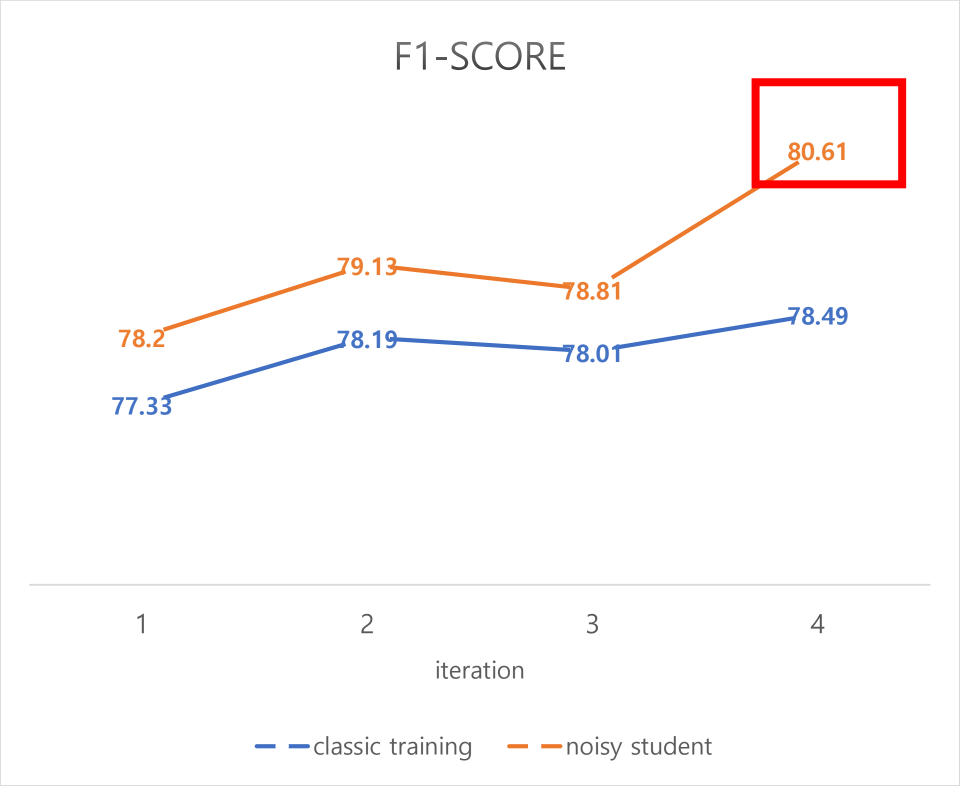
▷ F1-score



# 4. Test data change and F1-score measure



Pre-trained model: 78.85



Pre-trained model: 77.20

# 4. Test data change and F1-score measure

| | Positive | Negative |
|---|---|---|
| True | 2.8 | 1.4 |
| False | 0.37 | 5.4 |

$$\text{recall} = \frac{2.8}{2.8 + 5.4} = 0.34$$

$$\text{precision} = \frac{2.8}{2.8 + 0.37} = 0.88$$

06

# Conclusion

# Conclusion

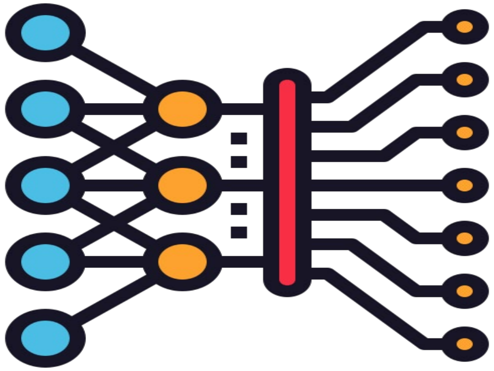
- ▷ Labeled data : dissimilar 200K < **similar** 2K
- ▷ Adding **noise** in Self-training can prevent learning incorrect pseudo labels
- ▷ The use of self-training is helpful in the **fine-tuning** stage.

# 후보 호감도 예측

긍정: 23.6%   부정: 76.4%   >   긍정: 20.8% 부정: 79.2%

# Self-training Effect

- Time – 27 hours (18K x 90m)
- Cost – 2,700,000 (18K x150)



- Artificial intelligence is used to avoid borrowing human hands, but data labeling consumes a lot of manpower
- Self-training can overcome the limits of deep learning
- Unseen data input the field can be utilized in real time

07

# Future works

# Future works

▷ **Self-training performance**

- Select confidence score
- KcBERT
- Experiment with large amounts of data
- Zoph, Barret, et al. "Rethinking pre-training and self-training." *Advances in neural information processing systems* 33 (2020): 3833-3845.

▷ **Multi-view Algorithm: Co-training**

- Several models work together to learn

# Q&A

Thank you