

Best Intelligent Group for Machine learning And Classification

---

# 2025 AIFFEL DLthon

: DKTC 분류 과제

# 01 팀원 소개

- **BIGMAC** Best Intelligent Group for Machine learning And Classification

팀장: 염철현



일반 대화 데이터 수집  
CNN, LSTM 모델 구현  
모델 스태킹 모듈화

팀원: 손병진



일반 대화 데이터 수집  
koELECTRA 모델 구현  
데이터 생성 및 가공

팀원: 김유은



일반 대화 데이터 수집  
KoBERT 모델 구현  
발표 자료 준비

팀원: 김천지



일반 대화 데이터 수집  
KoBERT 모델 구현  
텍스트 데이터 토큰 분석

# 목차

---

01 팀원 소개

---

02 일반 데이터 합성

---

03 개별 모델 구축 시도

---

04 데이터 전처리 및 증강

---

05 분류 과제 수행

---

06 결과 보고

---

## 02 일반데이터 합성

### • 데이터 확인

```
idx, class, conversation
0, 협박 대화, "지금 너 스스로를 죽여달라고 애원하는
아닙니다. 죄송합니다.
죽을 거면 혼자 죽지 우리까지 사건에 휘말리게 해?
정말 잘못했습니다.
너가 선택해. 너가 죽을래 네 가족을 죽여줄까.
죄송합니다. 정말 잘못했습니다.
너에게는 선택권이 없어. 선택 못한다면 너와 네 가족
선택 못하겠습니다. 한번만 도와주세요.
그냥 다 죽여버려야겠군. 이의 없지?
제발 도와주세요."
```

Training Data

협박	896
갈취	981
직장 내 괴롭힘	979
기타 괴롭힘	1,094
일반	-

Test Data(추정)

협박	~100
갈취	~100
직장 내 괴롭힘	~100
기타 괴롭힘	~100
일반	~100

데이터셋 확인 결과 주어진 Train Data에는 '일반' 데이터가 존재하지 않음  
 일반대화 분류를 위해 데이터 수집이 필요

## 02 일반데이터 합성

### • 데이터 생성



perplexity



### 일반 대화 데이터셋 구축을 위해 각자 250개씩 총 1,000개의 대화셋을 프롬프팅해 생성

: 20개의 주제로 생성 (생활, 사회, 전문, 환경 등)

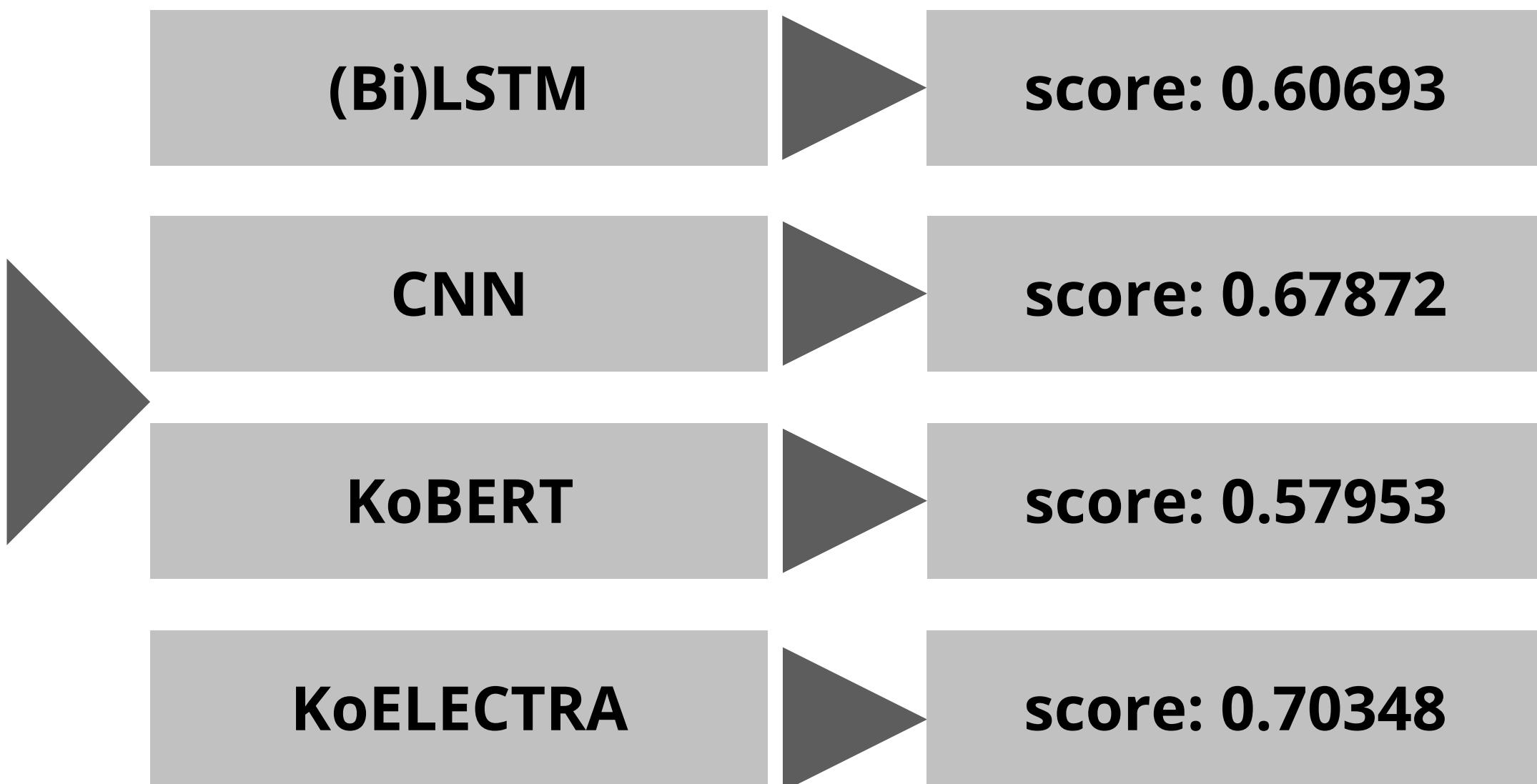
idx	class	conversation
1	일반	엄마, 이번 주말에 가족 여행 갈까요? 좋은 생각이네. 어디로 갈까? 제주도 어때요? 비행기 표가 비쌀 것 같은데. 그림 근로로 갈까? 가족은 어때요? 좋아, 속도 알아보고 맷집도 찾아보자. 알겠어요, 제가 찾아볼게요. 그래, 고마워. 이번 여행 정말 기대되네.
2	일반	아빠, 주말에 시간 있으세요? 품, 예 그레어? 꽃이 낚시 가고 싶어서요. 오, 좋은 생각이구나. 어디로 갈까? 청봉호 어때요? 좋아, 정夯는 내가 준비할게. 미끼는 제가 살게요. 그래, 고마워. 오랜만에 아빠랑 둘이 가네, 네, 정말 기대해!
3	일반	누나, 요즘 어떻게 지내? 베에서 친구들이 재워. 너는? 나도 학교를 알바로 빠빠. 가족들이랑 시간 보내네가 활동지? 말야, 다음 각자 브런치라고. 이번 주말에 다 같이 영화 보는 건 어때? 좋은 생각이다. 뭐 볼까? 액션 영화 어때? 좋아, 예매는 내가 할게. 간식은?
4	일반	할머니, 오사 건강은 어때세요? 허리가 좀 아프구나. 병원에 가보셨어요? 아니, 관찰을 거야. 제가 모시고 갈까요? 그래주면 고맙지. 내일 아침에 출근해. 고마워. 걱진 끝나고 맛있는 거 먹으려 가요. 그래, 꽃구나. 할머니랑 오랜만에 데이트네요. 그래게, 할미?
5	일반	형, 이번에 출전했더니? 축구해 고마워. 네 덕분이야. 내가 뭘 했다고. 항상 응원해줘서 형이 높아. 당연하지. 형 노력한 걸 냐니 잘 알잖아. 고마워. 이번 주말에 시간 돼? 응, 뭐? 험악 쏘고 싶어서. 오, 좋지 미리서 만날까? 신사동 빛칠 어때? 좋아, 기대된다.
6	일반	언니, 이번 주말에 결혼식이자? 음, 그레. 결혼되니? 많이. 도와줄 일 있어? 딜먼하趾. 누가 필요해? 둘러리 드레스 고르는 거 도와줄래? 그래, 기꺼이. 언제 갈까? 내일 저녁은 어때? 좋아, 7시에 만나자. 고마워. 든든하다. 당연하지. 언니 결혼식, 정말 기대되네.
7	일반	마들, 학교생활은 어때? 재밌어요, 아빠. 친구들은 잘 사귀고 있어? 네, 좋은 친구들 많아 사귀었어요. 다행이구나. 공부하는 데 어려움은 없어? 고마워요. 아빠, 그래. 오늘 저녁에 같이 공부해보자. 네,
8	일반	여보, 오늘 결혼기념일이야. 기념하고 있었어? 딜먼하趾. 경찰! 고마워. 너는 달신 선물 준비했어. 오, 봄때... 저녁에 집에 가서 열어봐. 알았어, 기대되네. 오늘 저녁 와 먹을까? 레스토를 예약해놨어. 와, 멋진데 사랑해. 여보. 너는 사랑해.
9	일반	이모, 오랜만이에요 잘 지내셨어요? 그래, 너는 어떻게 지내니? 대학 생활 재밌어요, 그렇구나. 학과는 학생은 맞고? 네, 재밌어요. 근데 과제가 많아서 힘들어요. 그래도 열심히 하고 있구나. 방학 때 우리 집에 놀러올래? 정말요? 좋아요! 언제 갈까요? 다음 달
10	일반	오빠, 생일 축하해! 고마워. 어, 선물도 준비했네? 당연하지. 열어봐. 와, 내가 갖고 싶어 하던 시계잖아. 기획하고 있었구나. 고마워. 마음에 들어? 너무 좋아. 오늘 저녁에 시간 돼? 응, 뭐? 맛있는 거 사줄게. 오, 좋지? 미리로 길까? 네가 좋아하는 곳으로 가자.
11	일반	너 이번 주말에 뭐해? 가족 모임이 있어서 못 나갈 것 같아. 아, 그렇구나. 자주 모이는 땅이야? 한 달에 한 번 정도? 좋겠다. 우리 집은 다음 달 바빠서 오기기 힘들어.
12	일반	친구애, 요새 표정이 안 좋마 보이던데 무슨 일 있어? 부모님랑 좀 디쳤어. 뭐? 무슨 일로? 친구들로? 서로 의견이 안 맞아서. 힘들겠다. 그래서 힘들겠다. 우리 표정이 좋을까?
13	일반	선네, 이번 주말에 고향 가족들과 만나러 가요. 좋으시겠어요? 선물은 준비하셨어요? 아직이요. 워가 좋을까요? 부모님께는 건강식들이 좋을 것 같아요.
14	일반	동이야 회식 날짜 잡았어? 아직. 다음 가족 일정 때문에 시간 맞추기가 힘들더라고. 그렇구나. 주말보다는 꽤일 저녁은 어때? 그것도 좋은 생각이네. 한동물이불기.
15	일반	마, 너 요새 살 좀 편 것 같는데? 아빠, 할머니가 몸보신이라고 매일 보양식을 해주서. 부럽다. 우리 할머니는 멀리 사서 저주 못 봐. 그래, 가족이 가까이 있다는 게 좋은 점도 있지.
16	일반	이번 방학에 뭐할 거야? 가족여행 갈 거야. 어디로? 제주도. 좋겠다. 우리 가족은 다음 시간 맞추기 때문에 여행을 못 가. 그렇구나. 시간 내서 대내외. 가족이랑 보내는 시간이 소중하지.
17	일반	너 요새 이렇게 피곤해 보여? 집에서 동생 둘보느라 힘들어. 부모님이 빛발이어서 말이야. 힘들겠다. 그래도 가족을 위해 노력하는 모습이 멋져.
18	일반	주말에 꾸愀해서 가족들이랑 통산 다녀왔어. 와, 좋았겠다. 다음 시간 맞추기 힘들지 않아? 그래도 한 달에 한 번은 꼭 시간 내려고 노력해. 가족 간의 유대감을 위해서.
19	일반	너 요새 표정이 밝아 보여. 무슨 좋은 일 있어? 음, 오랜만에 계신 언니가 오셨거든. 오, 그래도 가족이 더 모이니까 좋겠지. 말야, 오랜만에 가족이 다 모여서 좋겠다. 그래서 짱을 행복해.
20	일반	이번 주말에 위해? 풍생 결혼식이야. 축하해 둘러리 서나? 응, 처음으로 가족 결혼식에서 둘러리 서. 긴장되겠다. 그래도 특별한 경험이 될 거야.
21	일반	친구애, 요즘 연예는 어떻게 돼가? 별로 잘 안 풀려. 무슨 일 있어? 서로 바빠서 만날 시간이 없어. 그렇구나. 데일리 계획은 세워놨어? 주말에 시간 내서 하려고. 좋은 생각이야. 어디로 갈 거야? 한강공원에 피크닉 가려고. 남면적하네. 제일개 대내외.
22	일반	선네, 이번 프로젝트 어떻게 진행되고 있나요? 생각보다 어려울까 많아. 무엇이 가장 힘든가요? 팀원들과의 의견 조율이 쉽지 않아. 그렇군요. 제가 도와드릴 일이 있을까요? 회의 참석을 좀 도와줄 수 있어? 물론이죠. 언제가 좋으세요? 내일 오후 어때? 알겠습
23	일반	동마리 회장님, 다음 행사 준비는 잘 되고 있나요? 응. 대부분 끝났어. 도움이 필요한 부분이 있나요? 사실 홍보 쪽이 좀 부족해. 제가 SNS 홍보를 맡아볼까요? 정답? 그런 고맙지. 어떤 방식으로 할 건데? 인스타그램이랑 페이스북에 광고 올릴게요. 좋아, 기다
24	일반	선생님, 제 진로에 대한 상담 받고 싶어요. 그래, 어떤 고민이 있나? 문제와 이과 중 어느쪽으로 가야 할지 모르겠어요. 내 적성과 흥미는 어느 쪽에 더 가까워? 고쓰기를 좋아하지만 수학도 재밌어요. 그렇구나. 진로작성검사를 한번 해보는 게 어떨까요? 네, 좋아
25	일반	후배야, 취업 준비할 때 되고 있어? 아직 많이 부족해요. 선배님, 특별히 어려운 편이 있나? 자소서 쓰는 게 너무 어렵네요. 그렇구나. 내가 청식 총 해줄까? 정답요? 감사합니다. 언제 시간 빙정드세요? 이번 주말에 카페에서 만나자. 네, 정말 감사합니다. 많이!
26	일반	팀장님, 이번 분기 실적이 좋네요. 축하드립니다. 고마워요. 디 어려운 덕분이에요. 저희가 뭘 했거든요. 팀장님 리더십 덕분이죠. 과연이에요. 이번에 회식하는 게 어떨까요? 좋은 생각이네요. 언제가 좋을까요? 다음 주 금요일은 어떨까요? 좋습니다. 제가 장소
27	일반	풀레이트야, 우리 기숙사 규칙 좀 정해볼까? 그래, 어떤 걸 정하고 싶어? 청소를 쓰레기 버리는 날 정했으면 좋겠어. 좋아, 청소는 주말에 같이 하는 건 예쁜? 좋아. 쓰레기는 모일발로 나눠서 버리자. 그래, 평일은 내가 하고 주말은 내가 할까? 좋아. 그렇게 하지
28	일반	선배님, 출입하신 후 어떻게 지내세요? 회사 다니면서 책을 좋아해. 좋은 책이었는데요? 책을 점이 많아서 좋아. 흰 점 없으세요? 엄두망이 많아서 좀 버거워. 그렇군요. 어떻게 극복하고 계세요? 시간 관리를 철저히 하려고 노력 중이야. 저도 참고해야겠어
29	일반	친구애, 요즘 우울해 보이던데 무슨 일 있어? 가족들이랑 자주 대화해서 그래. 무슨 일인지 알해줄 수 있어? 내 장애 문제로 의견이 안 맞아. 많이 힘들겠다. 어떻게 해결하고 싶어? 대화로 물고 싶은데 쉽지 않아. 내가 중재해줄까? 그래 좋래? 고마워. 이번 주말에
30	일반	미주마니, 만녕하세요. 미시 온 지 얼마 안 됐죠? 네, 한 달 됐어요. 책을 잘 되고 계신가요? 네, 덕분에요. 혹시 불편한 점 있으면 언제든 말씀해주세요. 감사합니다. 그리고 보니 분리수거 요일을 잘 모르겠어요. 아, 매주 화요일이에요. 내일 마침에 현관 앞에

## 03 개별 모델 구축 시도

### • 모델 학습 및 성능 확인

- Train : Validation = 0.8 : 0.2

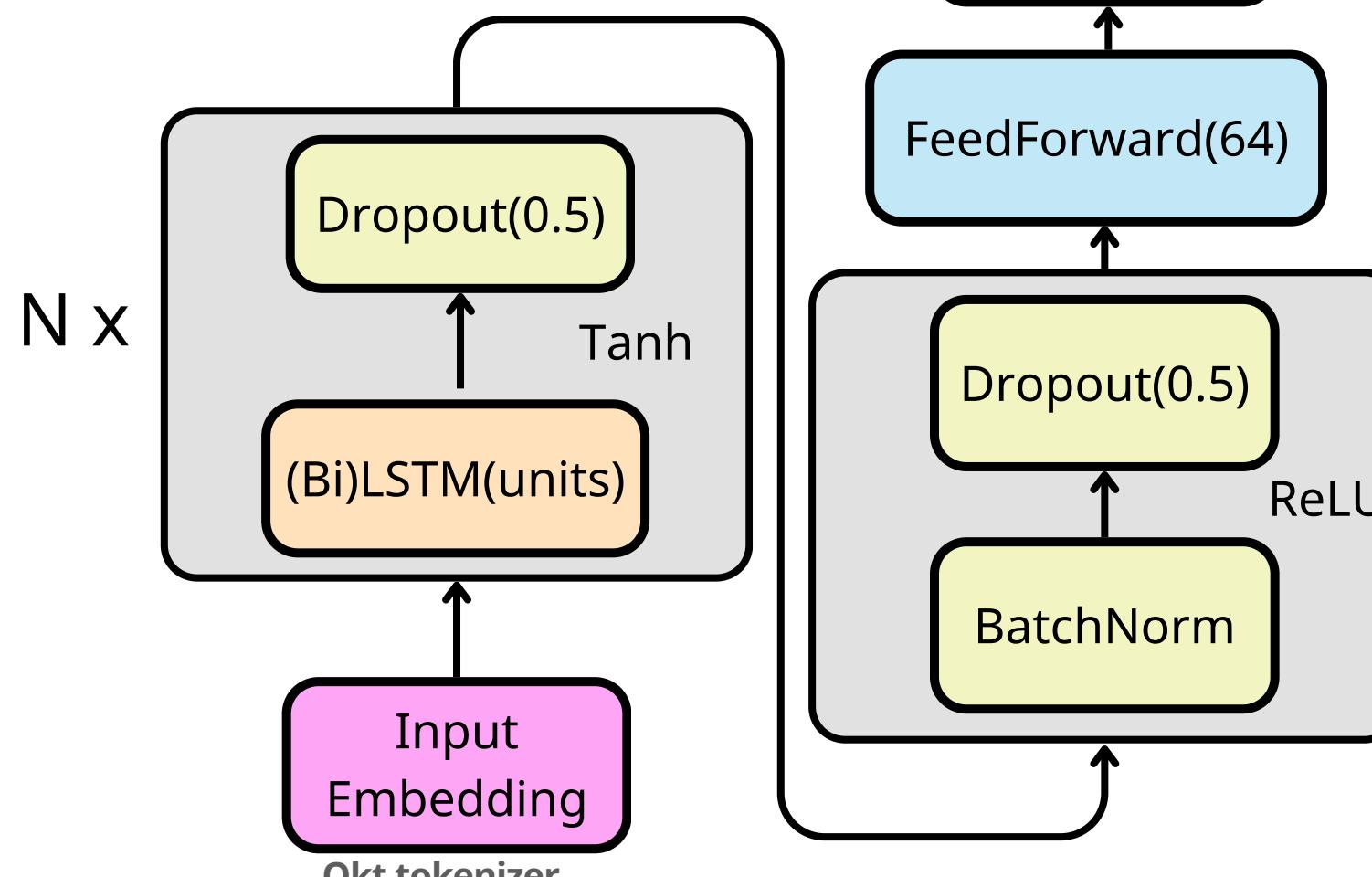
협박	896
갈취	981
직장 내 괴롭힘	979
기타 괴롭힘	1,094
일반	1,000



# 03 개별 모델 구축 시도

## 1) LSTM 모델

훈련 데이터: Train + 합성데이터



이후 BiLSTM 으로 모델 변경

### Hyperparameters

LSTM units(N=3)	128, 64, 32
최대 토큰 길이 *truncate 적용	150
임베딩 차원	128
단어 개수	10,000
옵티마이저	Adam

### Test 예측 클래스 분포

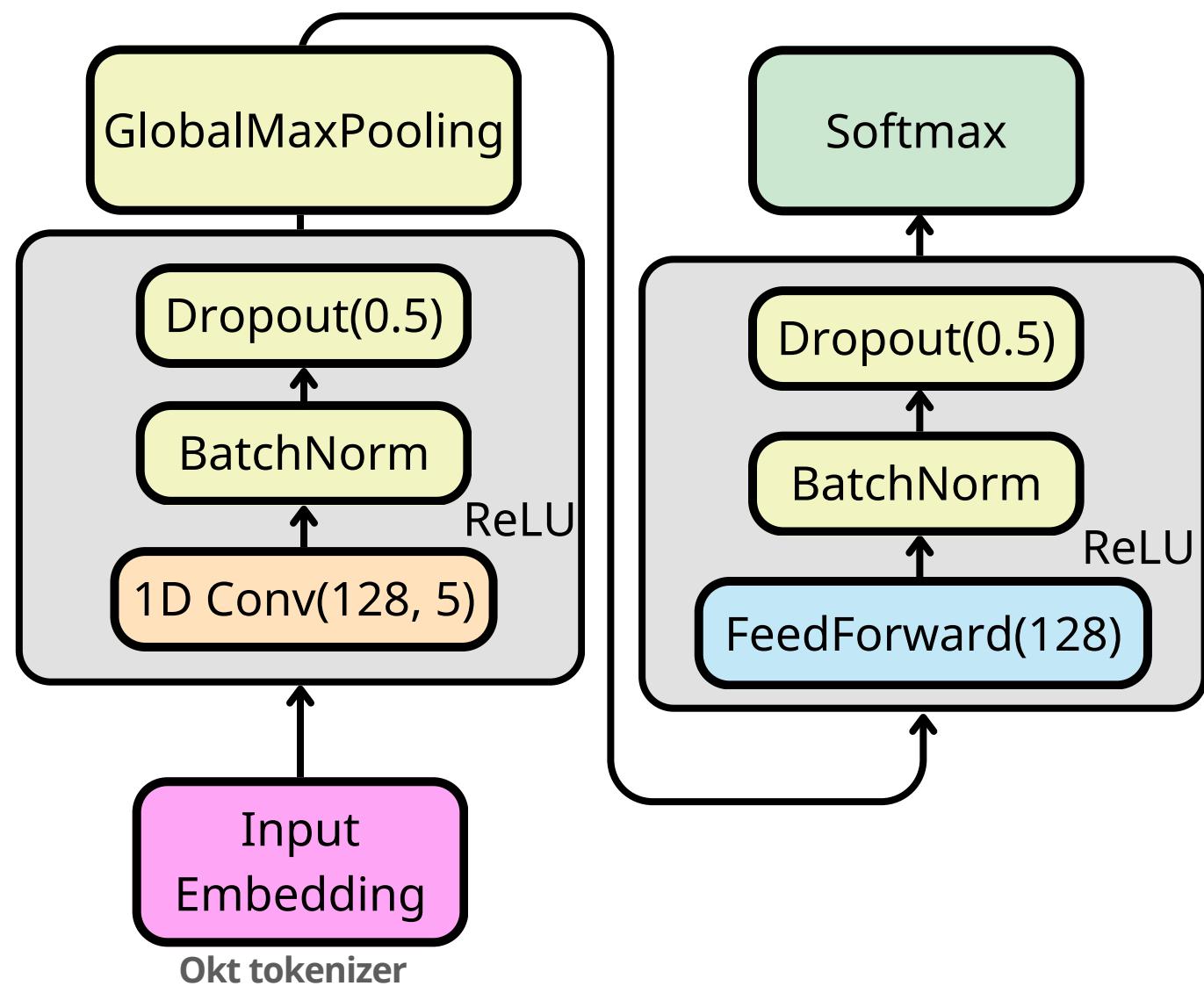
협박	80
갈취	115
직장 내 괴롭힘	133
기타 괴롭힘	156
일반	16

submission score: 0.60693

# 03 개별 모델 구축 시도

## 2) 1D CNN 모델

훈련 데이터: Train + 합성데이터



Hyperparameters

Regularizer	L2
최대 토큰 길이 <small>*truncate 적용</small>	150
임베딩 차원	256
단어 개수	10,000
옵티마이저	Adam

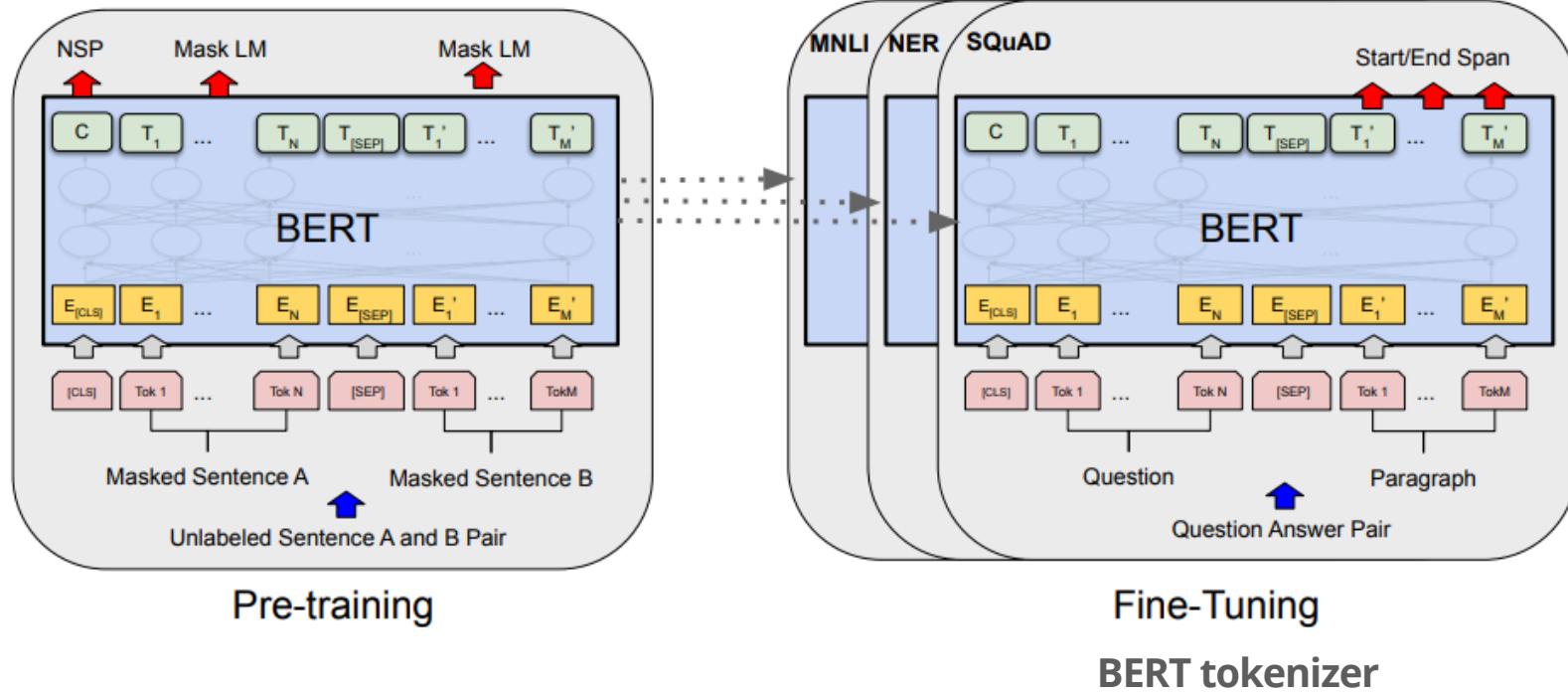
Test 예측 클래스 분포

협박	93
갈취	110
직장 내 괴롭힘	124
기타 괴롭힘	157
일반	16

submission score: 0.67872

# 03 개별 모델 구축 시도

## 3) KoBERT 모델



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

### Hyperparameters

Hidden Size	768
hidden_layers	12
Dropout_Prob	0.1
learning_rate	0.00002
옵티마이저	AdamW

### Test 예측 클래스 분포

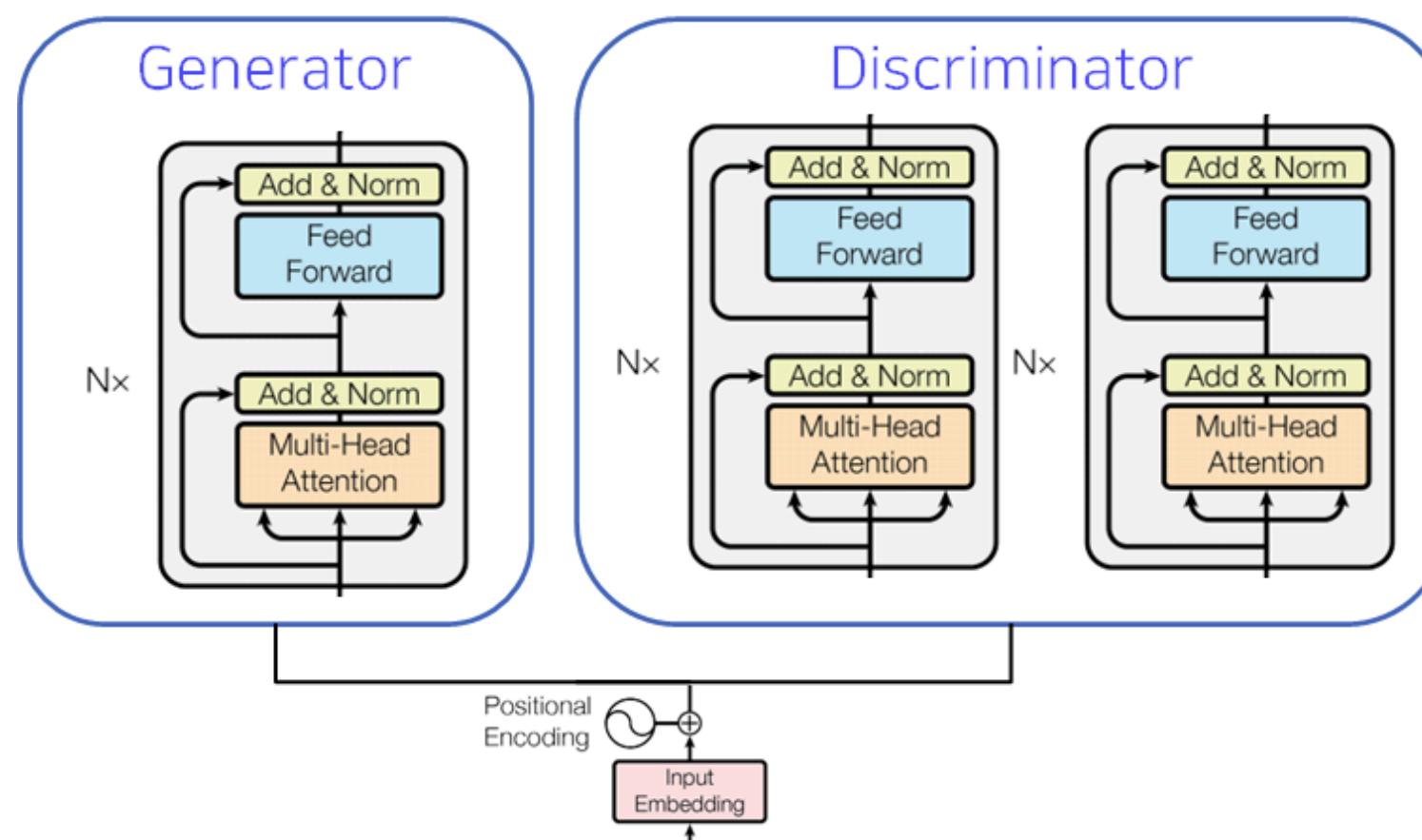
협박	156
갈취	72
직장 내 괴롭힘	117
기타 괴롭힘	113
일반	42

submission score:

0.57953, 0.42652

# 03 개별 모델 구축 시도

## 4) KoELECTRA 모델



**Hyperparameters**

hidden_layers	12
hidden_size	256
dropout_prob	0.1
max_length	200
옵티마이저	Adam

**Test 예측 클래스 분포**

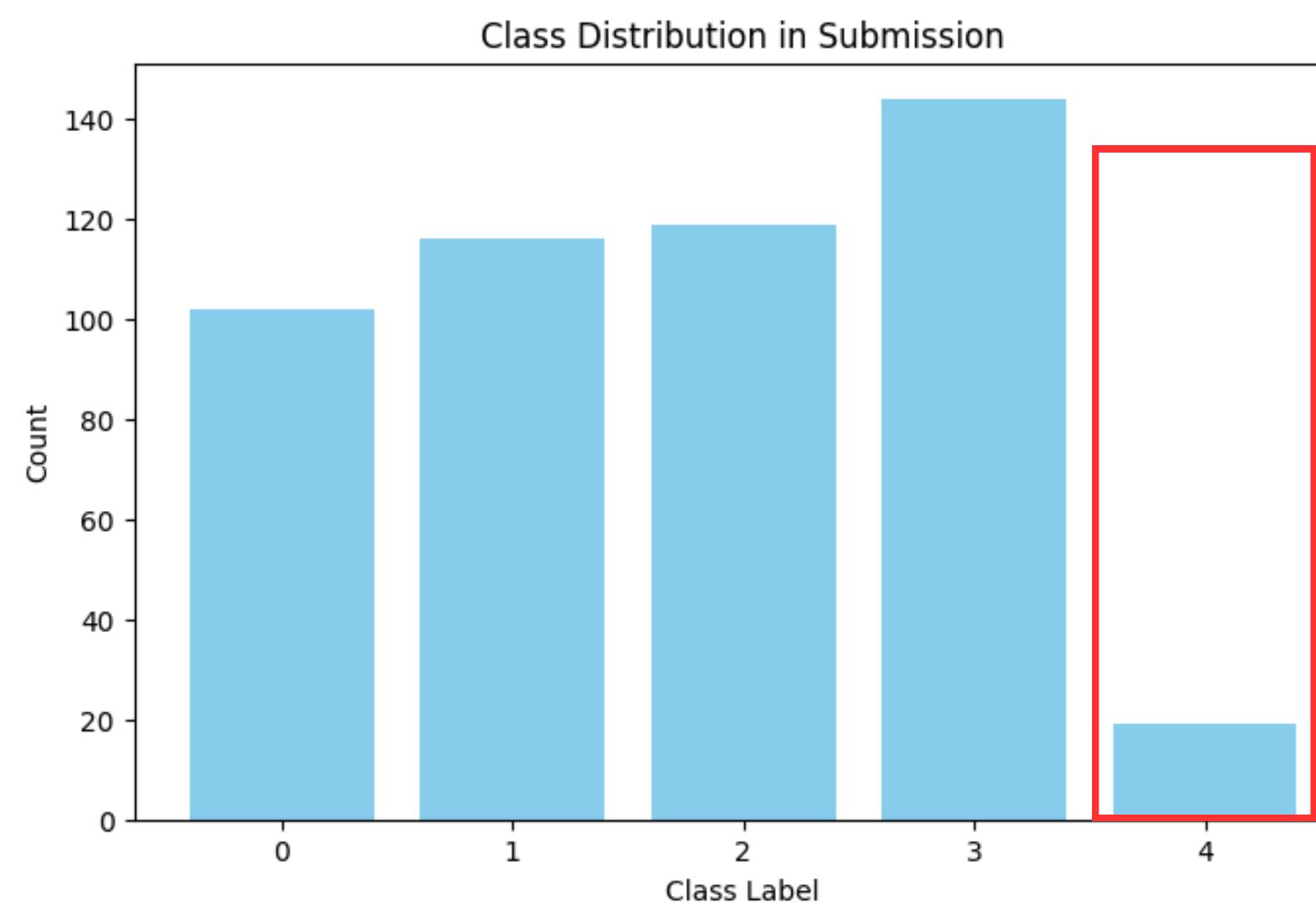
협박	102
갈취	116
직장 내 괴롭힘	119
기타 괴롭힘	144
일반	19

**submission score: 0.70348**

ELECTRA / Pre-training Text Encoders as Discriminators Rather Than Generators

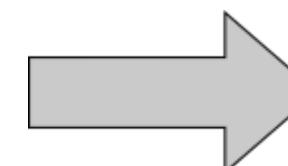
# 03 개별 모델 구축 시도

## 전반적인 문제점 발견:



## Validation Class Report

	precision	recall	f1-score	support
협박	0.82	0.91	0.86	80
갈취	0.83	0.96	0.89	94
직장 내 괴롭힘	0.94	0.91	0.93	90
기타 괴롭힘	0.92	0.76	0.83	113
일반	1.00	1.00	1.00	102
accuracy			0.90	479
macro avg	0.90	0.91	0.90	479
weighted avg	0.91	0.90	0.90	479

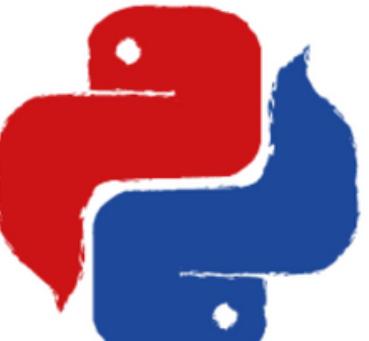


일반 대화 데이터셋 수정 필요!

# 04 데이터 전처리 및 증강

## kkma 토크나이저 기반 데이터 분석

- 각 클래스별로 자주 등장하는 토큰을 형태소 단위로 분석
- 분석된 토큰 중 문법적인 요소와 의미가 약한 요소를 불용어로 지정



KoNLPy

Kkma Class

### 형태소 분석

55. 느, 어, 약속, 하, 냐, 이번, 에, 약속, 하, 냐, 만큼, 못, 빼, 었, 더라, ?, ?, 부족, 하, 냐, 거, 이, 있, 잖아, ,, 그리하, 여도, 못, 빼, 냐, 거지, 돼지, 아, 어  
 66. 아줌마, 가, 사이좋, 게, 또, ?, 부른, 고, 보, 세요, 운전, 연수, 여기, 사기꾼, 이, 네, 아, 여기, 가, 일, 이, 시, 죠, 들어가, 다, 음주, 부터, 아줌마, 해, 요  
 67. 아, 아, !, 아줌마, 이월, 게, 하, 면, 어떻, 게, 하, 어요, ?, 이, 네, ?, 무슨, 월, 이, 시, 죠, ?, 아줌마, 가, 오토바이, 를, 치, 었, 잖아요, ,, 이렇게, 주차  
 68. 저, 어, 이번, 에, 육아, 휴직, 좀, 내, 려구요, 너, 는, 남, 자고, 하, 는, 데, 무슨, 줄, 육아, 휴직, 이야, 아내, 혼자, 너무, 힘들, 어, 하, 어서요, 그래서, 너  
 69. 절수, 도, 내가, 먼지, 내가, 만들, 냐, 같, 으니까, 큰일, 야, 현우, 야, 지, 있, 길래, 준비, 하, 고, 존, 제, 자체, 가, 알, 니, 작작, 하, 어, ., 하, 지, 마, 원  
 70. 야, 재, 형, 아, 아, 미치, 냐, 새끼, 가, 누구, 보, 고, 재, 형, 아래, 더럽, 네, 하, ., 하, ?, 야, 새끼, 야, 나, 가, 더, 기분, 나쁘, 아, 존, 제, 자체, 가, 새끼  
 71. 야, 저, 의, 민호, 보, 아, 바, 팔, 한, 짹, 밖, 에, 없, 어, 애, 들, 아, 아, 놀리, 지, 마, ., 야, 무엇, 을, 놀리, 냐다고, 그리, 나, 놀리, 는, 거, 아니, 야, 내, :  
 72. 빌리, 겠, 다고, 내노, 아, 가, 라, 알, 니, 야, 이거, 다, 안되, 는데, 둘려주, 세요, 되돌리, 어, 놓, 고, 뺄리, 안되, 어요, 제발, 요, 하, 르, 때, 진짜, 이, 새  
 73. 아, 좋, 은, 말, 로, 하, 르, 빼, 고, 가, 라, 아빠, 가, 사, 아, 주신, 거, 예, 요, 알, 니, 그냥, 빌리, 겠, 다고, 뺄리, 벗, 고, 가, 아, 아, 진짜, 안, 되, :  
 74. 야, 너, 가, 게임, 팀, 키, 름, 잠인, 맞, 냐, ?, 응, 맞, 는데, 워, ., 웨, 블르, 었, 는데, 키, 도, 조그마, 하, 이서, 초등학생, 도, 못, 이기, 르, 덩치, 구, 만  
 75. 하, 어도, 종, 에, 지감, 불러오, 아, 라, 산다는, 그리하, 여, 주, 냐다고, 주, 르래, 따라오, 아, 잘, 돈, 워, 가, 마세, 여, 느, 어, ., 하나, 는, ?, 잘, 종, :  
 76. 야, 지나가, 는, 애, 가, 종, 잘, 살, 냐다며, ?, 예, ., 형님, 제가, 우리, 학교, 이서, 잘, 종, 에, 하나, 이, 냐니다, ., 어리, 버리, 하, 기, 는, 하, 어도, 하나  
 77. 최대, 리, 는, 이제, 살, 종, 빼, 야지, 그리하, 여, 사람, 이, 너무, 둔하, 어, 보이, 네, 살, 빼, 아지, 종, 빼, 는, 것, 이, 어, 어, 때, 운동, 을, 하, 고, 있, 습  
 78. !, !, !, 치즈, 를, !, 듣, 어, 가있, 고, 드리, 끼까요, 죄송, 하, 냐니다, !, !, 주, 면, !, !, !, 저희, 다른난말, 이야, 워, 를, ., 고객, 님, 많, 아, 빼, 고,  
 79. 새우, 버거, 저희, 가게, 는, 쉬, 림프, 치즈, 버거, 있, 는데, 그, 것, 으로, 알, 니, 말, 이, 그렇게, 많, 아, !, 아, 아, 죄송, 하, 냐니다, ., 다른, 아서, 확인  
 80. 야, 손가락, 장애, 워, ?, ?, 귀도, 장애, 있, 냐, ?, 웨, 못, 알아듣, 어, 너, 지금, 워, 이, 라, 하, 었, 어, 웨, ?, 장애, 보고, 장애, 라고, 하, 는, 것, 이, 잘  
 81. 것, 이, 냐데, 하, 려고, 진짜, 나도, ?, ?, 안녕, 워, 라고, 엄마, 카, 톡, 방, 게, ?, 있, 지, 않, 을, 거, 야, 이런, 차리, 었, 나, 보, 지, 지랄, 이, 아, !, 개, :  
 82. 철, 종, 아, 아, !, !, 느, 어, 왜, 다시, 초대, 하, 고, ?, ?, 지랄, 이, 라고, 하, 었, 냐, ?, ? 많이, 크, 었, 네, 이, 새끼, 가, 버려지, 같, 은, 것, 이, 아지, :  
 83. 저기, 요, 할아버지, 여기, 버스, 안이, 기도, 한데, 환기, 도, ., 절, 안되, 니까, 마스크, 즘, 쓰, 어, 주, 세요, ., 너, 가, 웬, 테, 나, 보고, 쓰, 라, 말, 아이  
 84. 메기, 작, 아, 메기, 꽈, 트이, 냐데, ?, 빽, 았, 다, 눈, 이, 댓, 글, 마세, 요, 너무, 워, 아, 무슨, 진짜, 워, 야, 다, 워, 토, 나오, ., 다, 얼구, 르, 죄, 로, 얼굴  
 85. 사진, 보, 아, 못생기, 있, 다, 댓, 글, 닿, 러지, 아, 마세, 요, 워, 얼굴, 토, 나오, ., 다, 사진, 워, 야, ?, 눈, 이, 너무, 작, 아, 다, 신고, 하, 르, 것, 이, 냐니다  
 86. 너, 돈, 많, 냐, ?, 돈, 종, 빌리, 자, ., 돈, 안, 깔, 을, 거잖, 아, ., 내, 가, 웨, 빌려주, 어야, 되, 어, ?, 아, 아, 이, 쪼, 자고, 하, 는, 하, 계, ., 그럼, 확, !  
 87. 안, 비싸, 니까, 이쁘, 다, ?, 어, 어, 아, 아, 느, 어, ?, 되, 겠, 다, 안, 비싸, 던데, 이, 거, ?, 별로, ?, 화장, 하, 었, 어, 좀, 니, 가, 아, 아, 새, 거사, 이, 모  
 88. 친구, 야, 어, 어, ?, 무슨, 일, 이, 야, ?, 느, 어, 오늘, 화장, 하, 었, 어, ?, 아, 아, 응, ., 잘, 먹, 었, 다, ?, 이거, 이번, 신상, 아니, 야, ?, 너무, 이쁘, 다  
 89. 어이, 아가씨, !, !, 저, 아세, 요, 이, 거, ., ?, 이쁘, 냐데, 어디, 살, 아, ?, 저리, 가, 세요, ., 아, 이, 비싸, 계, 글, 르지, 말, 고, ., 몸매, 죽이, 는데, ?  
 90. 않, 았, 나요, 육, 하, 어서, 되, 는데, 인재, 씨, 하, 었, 습니다, 어, 어, 휴, 만, 만, 해, 요, 똑바로, 자르, 아, 버리, 고, 회사, 가, 죄송, 하, 냐니다, 그때, 시  
 91. 인재, 씨, 만, 만, 하, 어요, ?, 그, 계, 무슨, 말씀, 이, 세요, 지난주, 예, 사직, 고려, 중, 이, 라고, 말하, 지, 그, 건, 면담, 하, 다가, 감정, 이, 옥, 하, 어서  
 92. 저기, 이, 요, ., 저, 좀, 도주, 세요, ., 무슨일, 이, 세요, ?, 도와, 드리, 끼게요, ., 다치, 시, 냐, 델, 냐, 없, 으세요, ?, 감사, 하, 냐니다, ., 저, 가, 지금  
 93. 웨, 콜라, 하나, 매점, 가, 아서, 어, 어, 내가, 거스름돈, 야, 택, 도, 없, 어, 웨, 코털, 갖고오, 아, 원, 어이, 뻥, 두개, 두개, 예, 색, 키, 야, 돈, 없, 어, 뒤집

### 불용어 지정

idx	class0	class1	class2	class3
0	어	.	.	.
1	.	어	하	어
2	하	?	?	?
3	?	이	이	하
4	이	아	어	아
5	아	하	아	이
6	는	고	는	는
7	가	가	가	가
8	고	는	고	야
9	!	돈	ㄴ	고
10	ㄹ	야	네	ㄴ
11	야	ㄴ	ㅂ니다	나
12	었	나	에	었
13	나	에	었	지
14	ㄴ	주	습니다	니
15	을	네	나	네
16	지	없	ㄹ	거
17	거	ㄹ	야	!
18	에	었	도	도
19	게	!	거	게

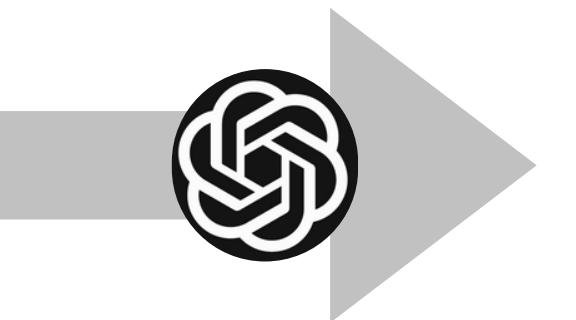
# 04 데이터 전처리 및 증강

## kkma 토크나이저 기반 데이터 분석

클래스별로 자주 등장하는 단어를 바탕으로 일반 대화 생성

### 클래스별 빈출 단어

100	ㄴ다	까지	과장	다고
101	발	미안	좋	좋
102	말하	가지	아요	ㄴ다
103	돈	싶	생각	으면
104	맞	드리	때	때
105	이야	잘	으면	자
106	정말	원	싶	엄마
107	서	엄마	한테	사
108	잘	싫	시간	듣
109	다고	마	라고	더
110	라고	이번	수	건
111	살리	때	자네	랑
112	건	죠	이번	ㅁ
113	이러	빨리	ㅁ	무엇
114	어떻	그럼	진짜	이거
115	여기	의	그냥	님
116	ㄹ지	수	업무	말하
117	새끼	일	자	고객
118	모르	ㄹ래	조	살
119	때	나오	의	치
120	미안	돌려주	이것	집
121	듣	그렇	휴가	애



### 일반 대화 생성

```

python
 dialogs = [
    "엄마한테 뭐라고 말해야 할까?", "그냥 솔직하게 이번 여행에 대해 말씀드려.", "하지만 엄마가 싫어할 것
    ["친구랑 이번 주말에 여행 가려고 해.", "진짜? 어디로 가?", "여기 근처 좋은 곳 알아봤어.", "엄마한테 한
    ["엄마가 이번 여행 반대하셔.", "왜? 시간도 있고 좋은 기회인데.", "위험할까 봐 걱정하시는 것 같아.", "친구들이랑 여행 갈 건데 같이 갈래?", "좋지! 어디로 가?", "아직 정하진 않았어. 시간 되면 같이 계획해
    ["요즘 너무 힘들어.", "왜? 무슨 일 있어?", "그냥 요즘 모든 게 어렵게 느껴져.", "엄마한테 고민 얘기해
    ["엄마가 내 결정에 대해 뭐라고 하셨어?", "이번에는 그냥 네가 선택하도록 두셨대.", "진짜? 기대도 안 했는데
    ["내 친구가 이번에 유학 가.", "와, 대단하다! 어디로?", "미국으로 가. 진짜 부러워.", "그리고, 나도 외국
    ["나 요즘 너무 바빠.", "무슨 일로?", "학교랑 학원 때문에 시간이 없어.", "그래도 가끔은 쉬어야 해.", "이번 주말에 뭐 할 거야?", "아직 계획 없어. 왜?", "여행 갈까 해서.", "좋아! 어디 가?", "아직 정하진
    ["이번 학기에 성적 올리고 싶어.", "그럼 공부 계획 세웠어?", "아직 어떻게 해야 할지 모르겠어.", "엄마
    ["언제까지 이 일을 끝낼 수 있을까요?", "과장님께서 내일까지 꼭 끝내라고 하셨어요.", "아, 그렇군요. 정답은
    ["엄마가 해주신 음식은 정말 맛있어요.", "맞아요! 특히 김치찌개가 정말 최고죠.", "저도 그 생각 했어요.
    ["돈을 많이 벌고 싶다는 생각이 들 때가 있어요.", "저도 그래요. 그런데 돈만으로 행복할 수 있을까요?", "발이 아파서 걷기가 힘들어요.", "어머, 괜찮아요? 무슨 일이 있었나요?", "어제 운동을 너무 많이 해서 그런가?
    ["이 프로젝트에 대해 어떻게 생각하세요?", "저는 좋은 아이디어라고 생각해요.", "저도 그렇게 생각했어요.
    ["요즘 정말 바쁘네요. 시간이 부족한 것 같아요.", "맞아요. 저도 일이 많아서 정신이 없어요.", "그래도 우리는
    ["이거 정말 맛있네요! 어디서 샀어요?", "엄마가 직접 만들어주셨어요.", "와, 정말 대단하시네요. 레시피를 알려주세요.", "요즘 무슨 책 읽고 있어요?", "'돈의 심리학'이 책이에요. 정말 흥미로워요.", "아, 저도 그 책 읽어보려고"
    ["오늘 점심 뭐 먹을까요?", "저는 김치찌개가 먹고 싶어요!", "아, 저도 좋아해요! 맛있을 것 같네요.", "그럼
]

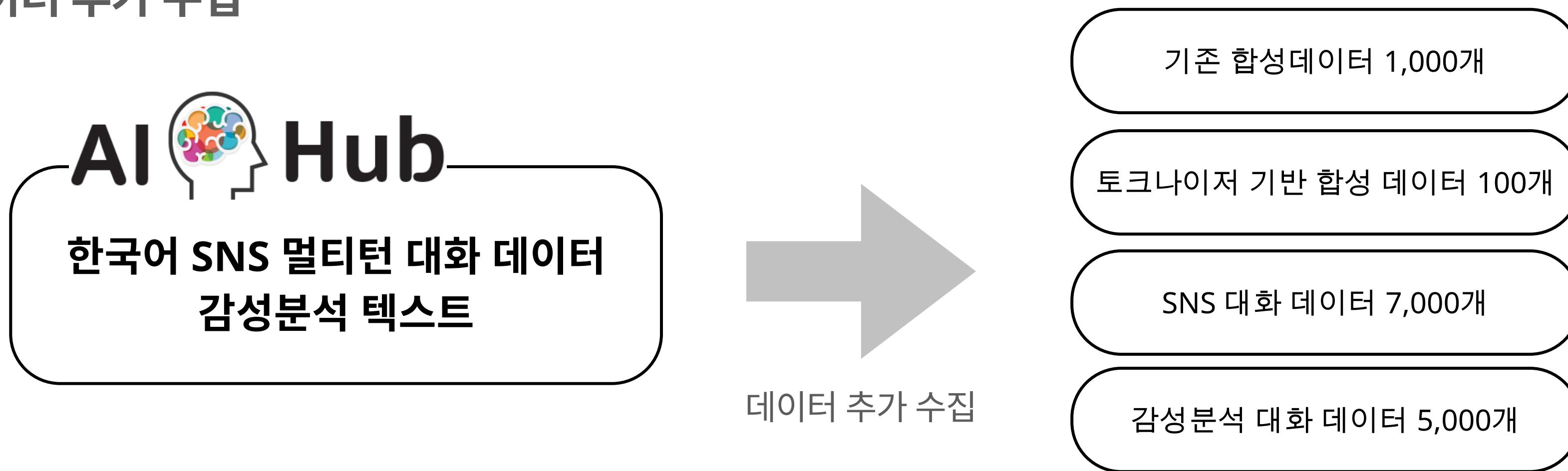
```

무엇이든 물어보세요

+ 검색 이성

## 04 데이터 전처리 및 증강

### 데이터 추가 수집

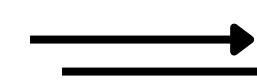


AI Hub에서 공개되어있는 SNS 대화 데이터와 감성분석 데이터 추가 수집  
일반대화 데이터셋 13,000여개로 확장

## 04 데이터 전처리 및 증강

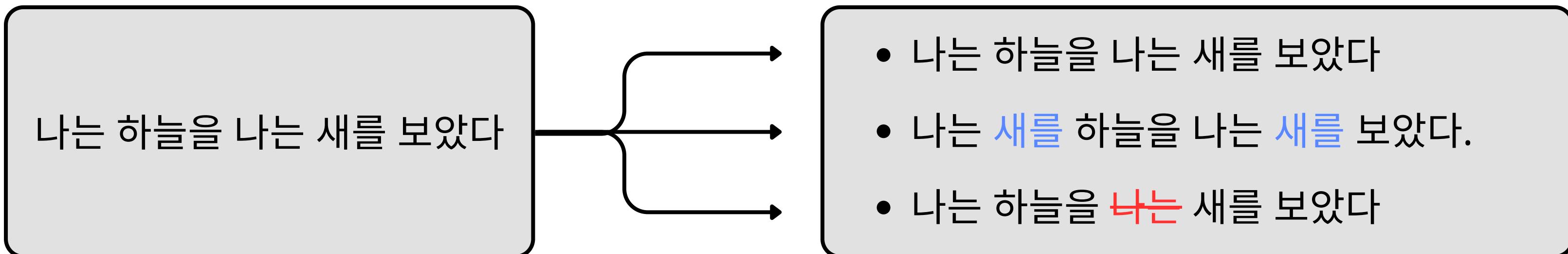
### 데이터 증강

- Random Shuffle (base)
- Random Insertion
- Random Deletion



3배 수의 데이터 확보

- Original Data (A)
- Randomly Insert (A1)
- Randomly Delete (A2)



# 05 분류 과제 수행

## 양상블 기법 도입 : 2단계 스태킹 (Homogeneous + Heterogeneous)

**Stage 1: 동형 양상블(Homogeneous Ensemble Stacking)**

간

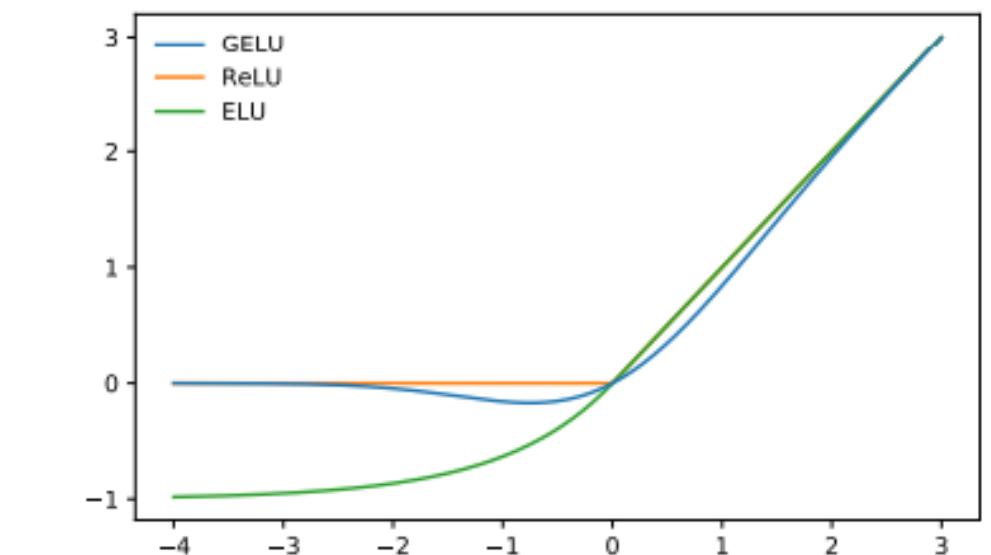
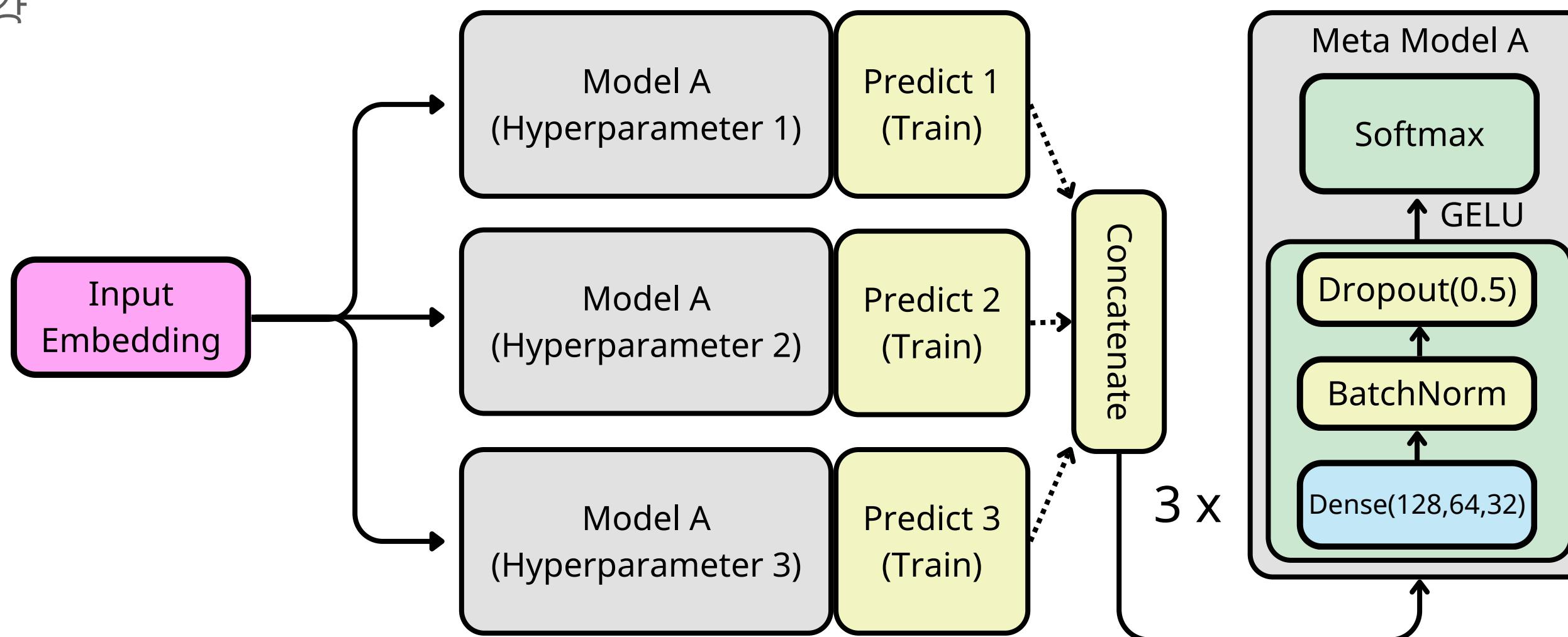


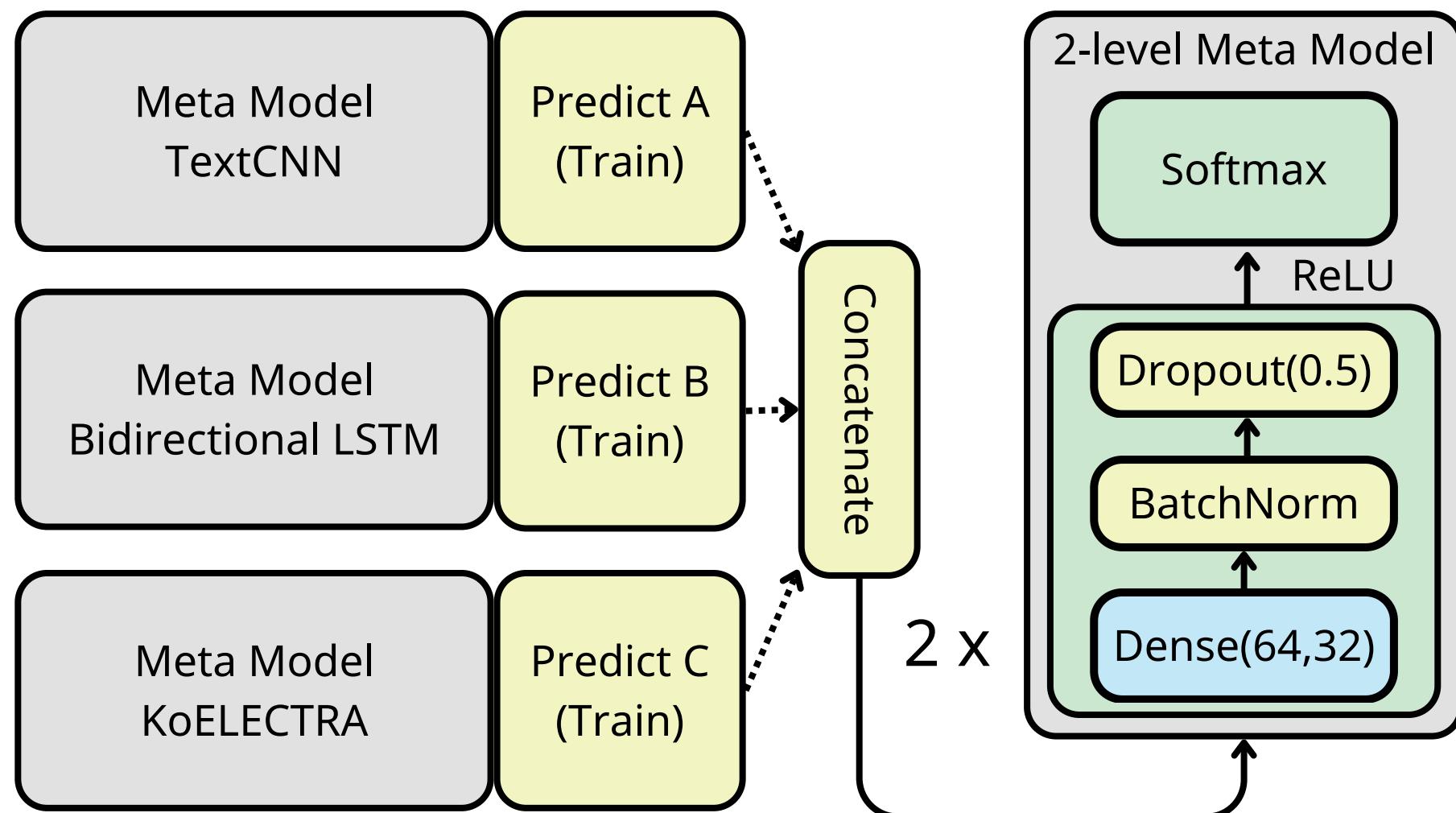
Figure 1: The GELU ( $\mu = 0, \sigma = 1$ ), ReLU, and ELU ( $\alpha = 1$ ).

# 05 문류 과제 수행

## 양상을 기법 도입 : 2단계 스태킹 (Homogeneous + Heterogeneous)

Stage 2: 이형 양상을(Heterogeneous Ensemble Stacking)

훈련 데이터: Train + 합성, 일반 데이터 추가 및 증강

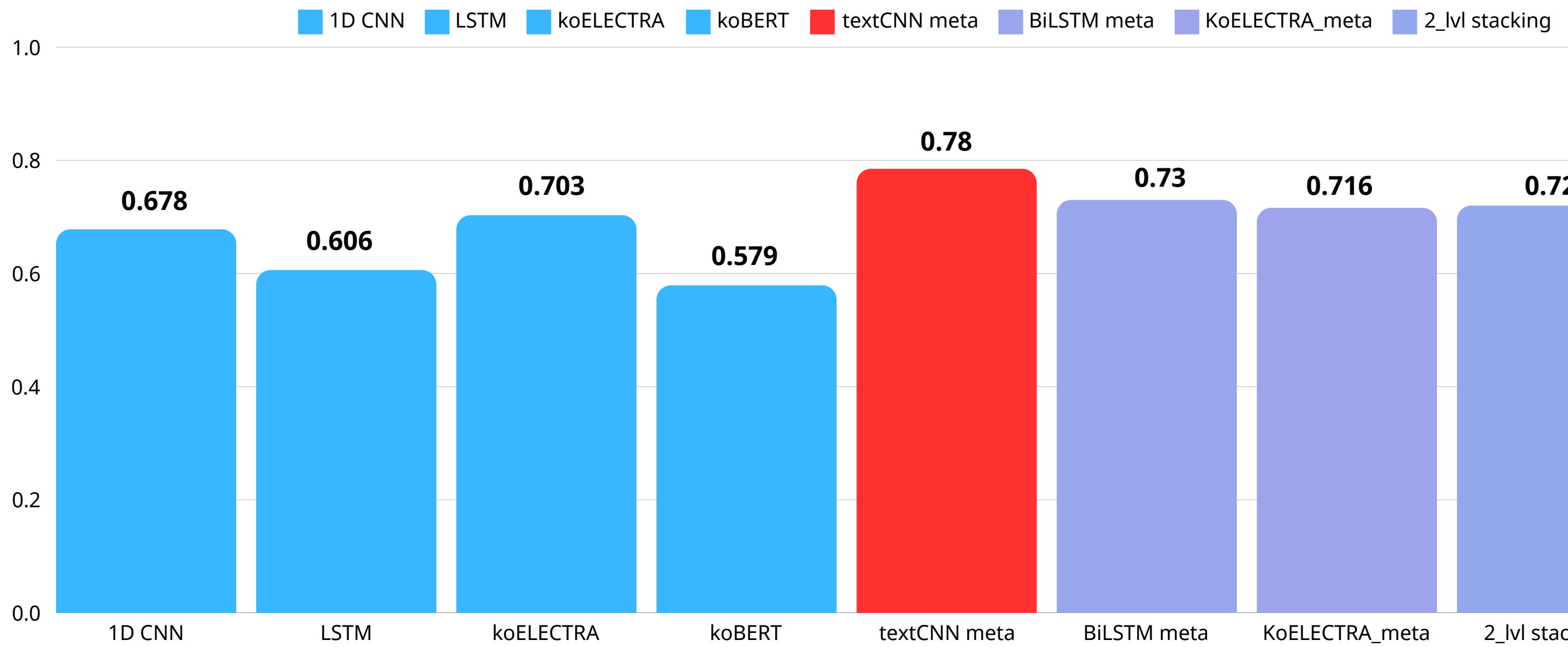


각 메타 모델 단계마다  
Test data에 대한 예측을 수행 가능!

메타 모델	Evaluation
TextCNN	0.78527
Bi-LSTM	0.73422
KoELECTRA	0.71634
2-level Meta	0.72628

## 05 분류 과제 수행

### 성능 확인 및 최종 모델 선정



# 06 결과 보고

## 최고 성능 모델: TextCNN 동형 양상을 스태킹

### 양상을 하이퍼파라미터 설정

L2 regularizer 적용

Model 1:  
Conv layer kernel : 5  
dropout = 0.5

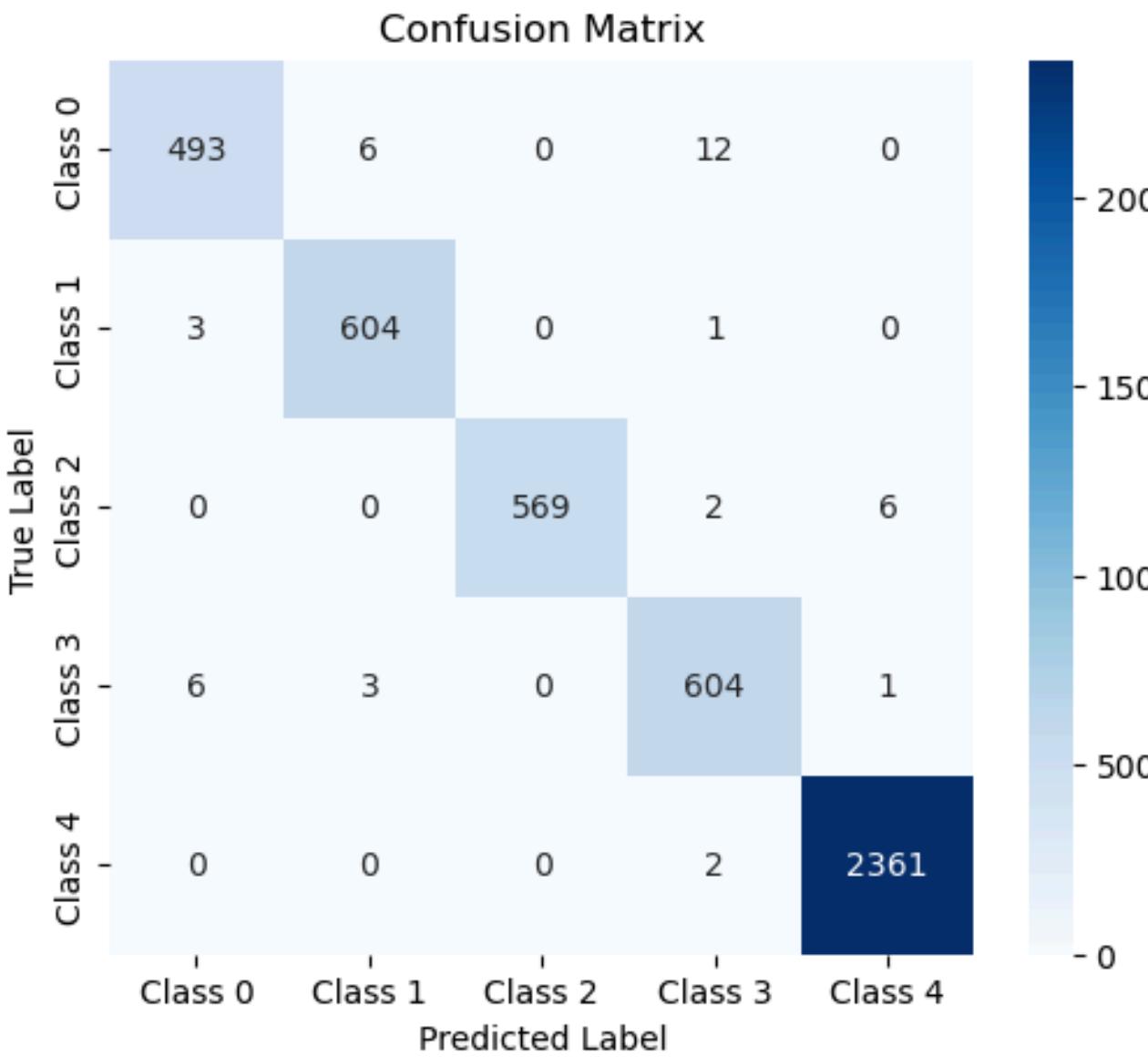
Model 2:  
Conv layer kernel : 4  
Dropout = 0.4

Model 3:  
Conv layer kernel : 3  
Dropout = 0.3

하이퍼파라미터 세팅이  
다양하지 못해 과적합 우려

Validation Set에 대한 검증

Validation Set에 대한 F1-score: 0.9910



### Test Data에 대한 예측

Test Data에 대한 F1-score: 0.78527

협박	103
갈취	113
직장 내 괴롭힘	115
기타 괴롭힘	108
일반	61

Validation data에 대한 과적합과  
일반 데이터 생성 및 가공 부족으로  
Test data에 대한 일반화 성능이 낮은 것을 관찰

**경청해주셔서 감사합니다.**

---

**QnA**