

Purifier L

Team. 사남자



본 발표에는 비속어가
다수 포함되어 있음을 알립니다



Index

01

온라인 공간에 범람하는 욕설

02

문맥 인식 비속어 필터 구현

03

깨끗한 공공 온라인 환경 조성



01

온라인 공간에 범람하는 욕설

01. 온라인 공간에 범람하는 욕설



[“이러다 진짜 죽을 거 같다” 류지혜, 악플 댓글 자제 호소](#)

헤럴드경제 | 1일 전 | 네이버뉴스 | [🔗](#)

[류지혜 SNS캡처] [헤럴드경제=모바일섹션] 전직 유명 프로게이머의 아이를 낙태했다고
취중 고백한 뒤 SNS를 통해 자살을 암시하는 듯한 글을 올렸던 레이싱 모델 겸 1인 BJ인 류



[배우 이상아 측 “딸 협박 등 악성 댓글 고소”](#)

연합뉴스 | 2019.01.08. | 네이버뉴스 | [🔗](#)

배우 이상아 [연합뉴스 자료사진] 배우 이상아가 도가 넘은 **악성 댓글**을 단 누리꾼들을 고
소했다. 소속사 마라톤엔터테인먼트는 8일 “이상아에 대한 끊임없는 허위사실 유포와 악성

[‘OO충’ 등 자살 부추기는 댓글 판친다](#)

매일경제 PICK | [📖](#) A5면1단 | 2018.12.02. | 네이버뉴스 | [🔗](#)

특정인을 향한 **악성댓글**과 신상털기를 비롯해 자살을 조장하는 콘텐츠들이 자정작용 없이 돌아다니며 시민들
을 극단적 선택으로 내몰고 있다는 분석이다. 방송통신심의위원회(방심위)에 따르면 올해 1월부터 10월...

‘응 못봤다 씨발개좆같은개보빨개씨발애미터진좆같은개년아 이제 됐냐? 그래 처음본다 개씨떨년들아’

‘니 애미도 10분에 5명을 못받던데 너는 탑에서 다섯번을 대주냐 이새끼야’

‘빨리좀 와라 이기야 왜 안오노... 노무 안온다 이기야’

‘니가 태생부터 더러운 잡종새끼라 그런거다’



변형

썬이벌 뽕럼프 새끼가 똤질라고



썬이벌 뽕럼프 **가 똤질라고

오탐

음식물 쓰레기는 따로 버려주세요



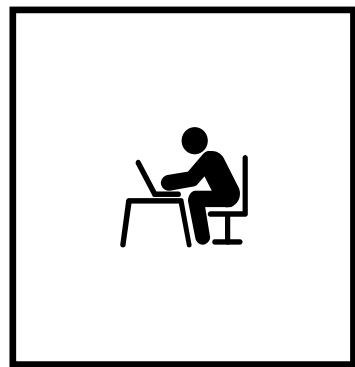
음식물 ***는 따로 버려주세요

맥락

니 머리에 든 뇌는 우동사리냐?

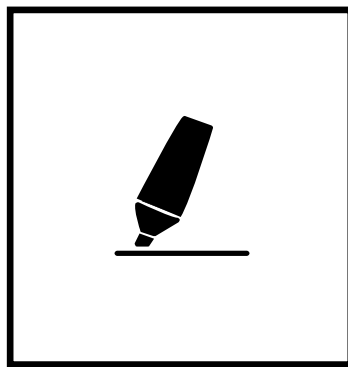


니 머리에 든 뇌는 우동사리냐?



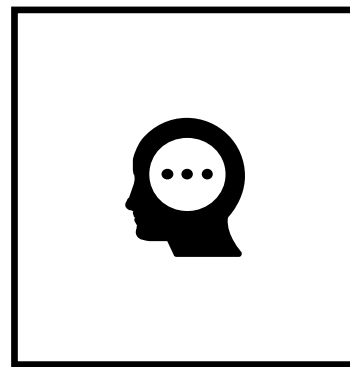
악플로 인한 피해

+



기존 모델 문제

+



표현의 자유 보장

새로운 **비속어 필터**의 필요성 대두



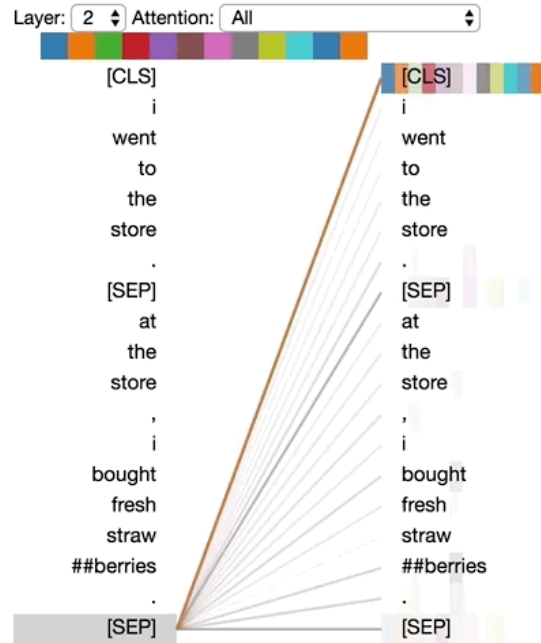
02

문맥 인식 비속어 필터 구현

System	Dev		Test	
	EM	F1	EM	F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
BERT _{BASE} (Single)	80.8	88.5	-	-
BERT _{LARGE} (Single)	84.1	90.9	-	-
BERT _{LARGE} (Ensemble)	85.8	91.8	-	-
BERT _{LARGE} (Sgl.+TriviaQA)	84.2	91.1	85.1	91.8
BERT _{LARGE} (Ens.+TriviaQA)	86.2	92.2	87.4	93.2

SQuAD Task에서 **Human lv**을 넘어선 NLP 모델

attention

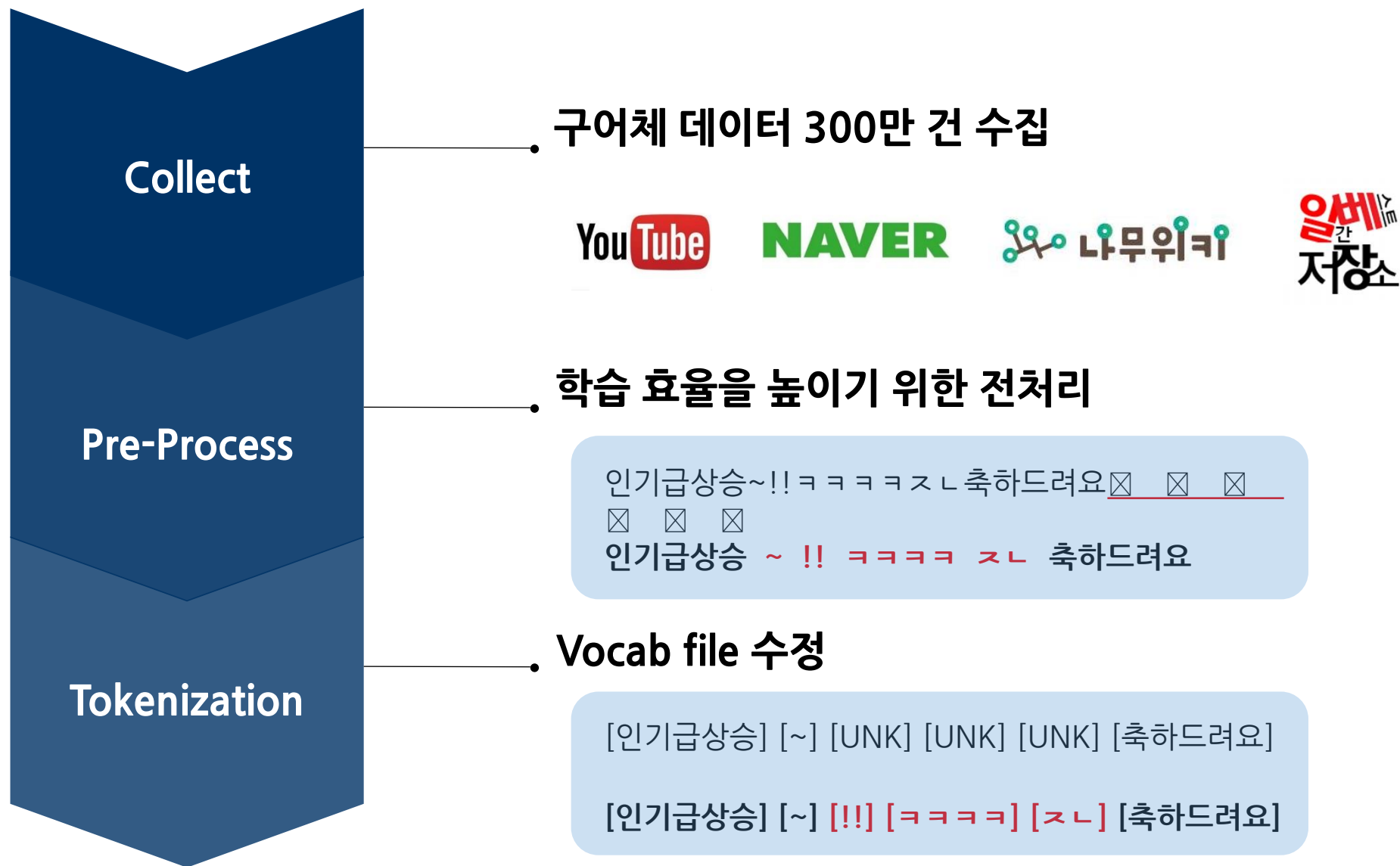



bidirectional

→ ←
새로운 아침이 밝았다

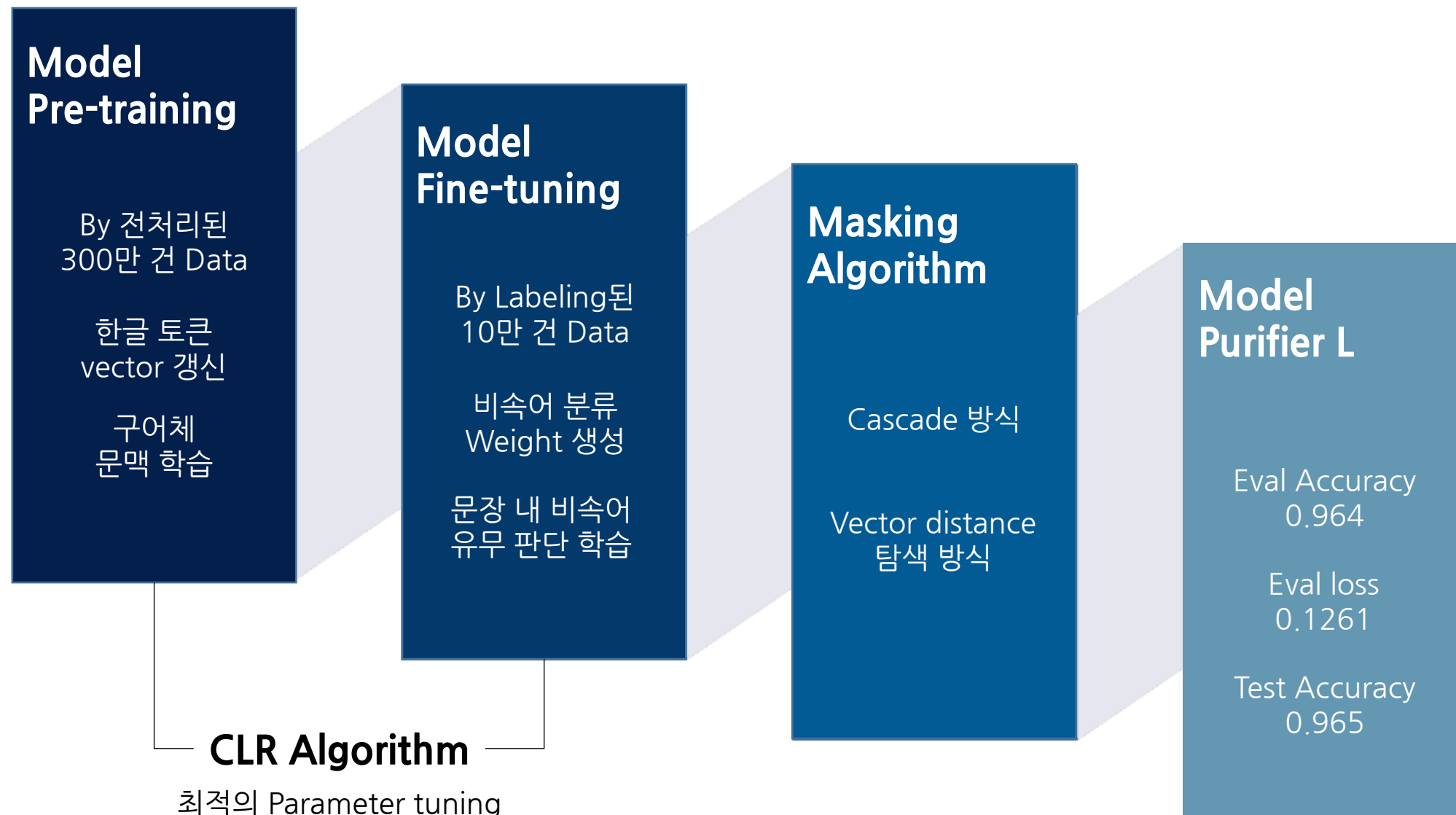
→ ←
맛있는 아침을 먹었다

두 가지 기법으로 글의 맥락을 학습



label	Comments	Bad words	Database
0	너네 아빠 조기축구회에서 패스 못 받음		
1	씨발 ㅋㅋㅋ 갑자기 훅 들어오넥ㅋㅋ	씨발	
0	니가 만든 모델 eval 이랑 test 50퍼 차이남		
1	그 동전 몇푼 벌어보겠다고 여기와서 주둥이를 털어대냐		
1	둘다 사고사 아님? 이게 좌좀들이 말하는 평등한 사회노?	좌좀	
1	개새끼 그딴 식으로 굴면서 사는건 유전이냐?	개새끼	
0	선생님, 왜 지금 검열로도 모잘라 더 만드려고 하십니까?		
1	해주면 해준다고 지랄 안해주면 안해준다 지랄... 어휴	지랄	

02. Purifier L 제작 & 예측 결과



썩이벌 뱅럼프 새끼가 똥질라고
감히 여왕폐하 면전에서 불경하게 **구노**

변형

*** 뱅럼프 ***
감히 여왕폐하 면전에서 불경하게 **

넌 진짜 **쓰레기**다
음식물 **쓰레기**는 따로 버려주세요

오탐

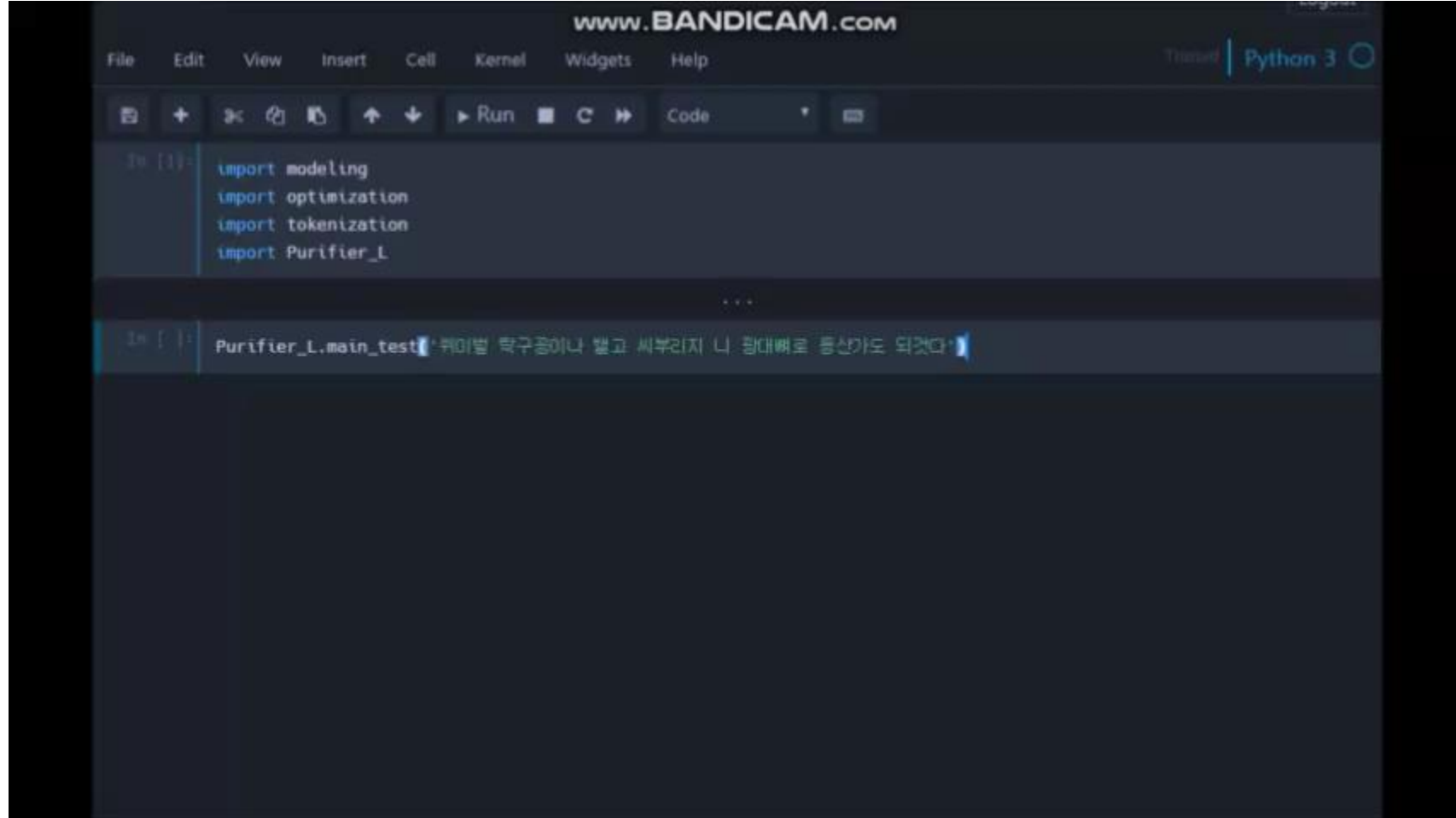
넌 진짜 ****
음식물 **쓰레기**는 따로 버려주세요

아줌마 **우동사리** 하나 추가요
니 머리에 든 뇌는 **우동사리**냐?

맥락

아줌마 **우동사리** 하나 추가요
* 머리에 든 뇌는 *****

02. 시연 영상



The screenshot shows a Jupyter Notebook interface with a dark theme. The top bar includes the website 'www.BANDICAM.com' and a menu with 'File', 'Edit', 'View', 'Insert', 'Cell', 'Kernel', 'Widgets', and 'Help'. On the right, it says 'Trust | Python 3'. Below the menu is a toolbar with icons for file operations, a 'Run' button, and a 'Code' dropdown. The notebook contains two code cells. The first cell, labeled 'In [1]:', contains the following Python code:

```
import modeling
import optimization
import tokenization
import Purifier_L
```

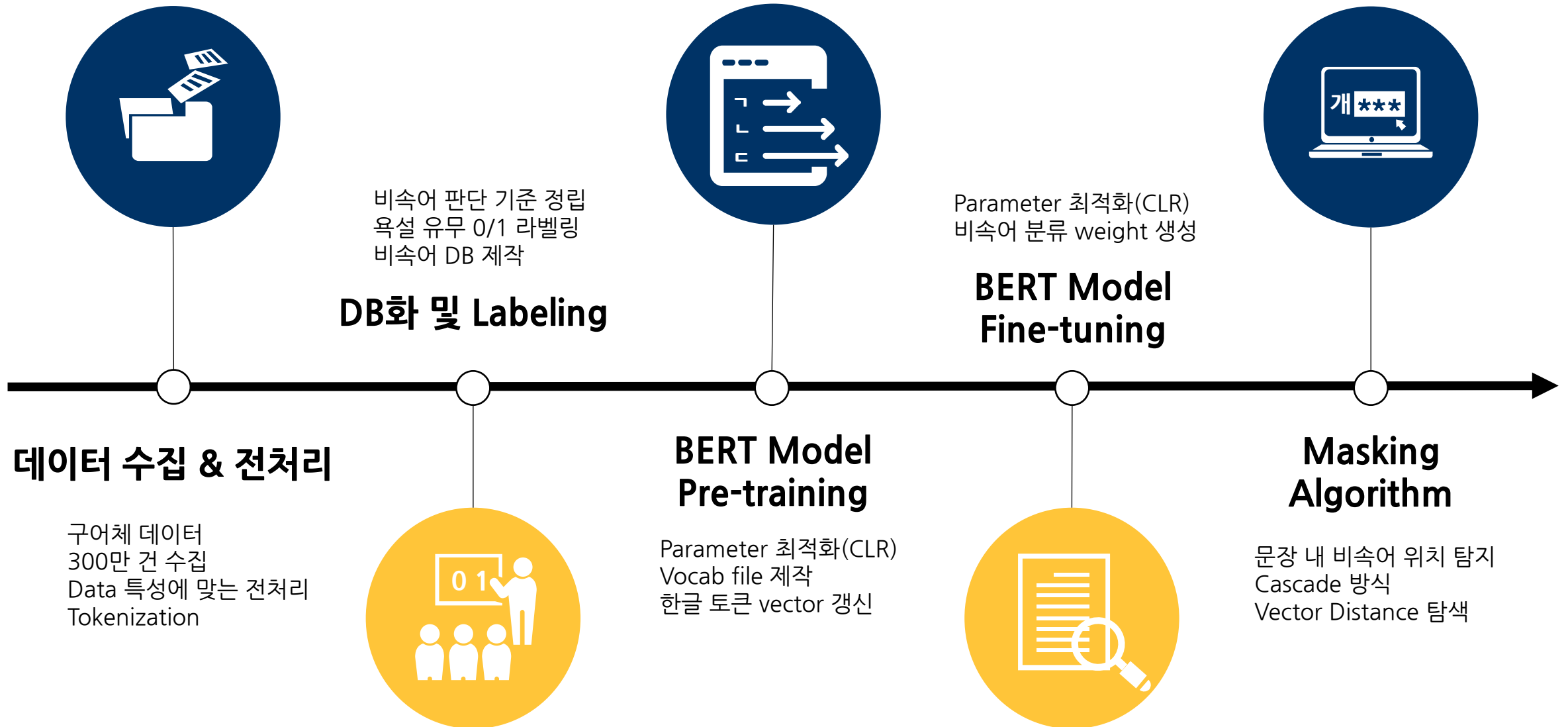
. The second cell, labeled 'In []:', contains the code

```
Purifier_L.main_test('썬이벌 탁구공에나 빨고 씨부리지 니 할대빠로 흥산가도 되겠다')
```

. The text in the code is a mix of Korean and English, with some characters appearing to be garbled or misspelled.

‘썬이벌 탁구공에나 빨고 씨부리지 니 할대빠로 흥산가도 되겠다’

02. Total process





03

깨끗한 공공 온라인 환경 조성



국가 민원 및 청원 게시판 적용

공공 서비스 질적 향상
합리적인 커뮤니케이션 유도

Purifier L

청소년 주요 방문 사이트 필터 적용

올바른 언어습관 함양
부정적 언어 노출 빈도 감소



Chrome Addon **사용자 선택형** 비속어 필터

표현의 자유 보장
악플로 인한 사회갈등 완화

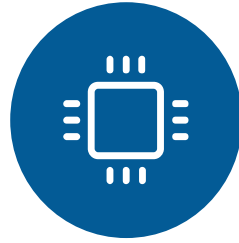


01

비속어 DB 구축

Purifier L 예측 결과
분석 DB를 구축

DB 기반 추가 Tuning



02

Masking Algorithm

시간 단축
정확도 개선



03

Chrome Addon

Purifier L 기반
크롬 확장 프로그램 구현

Q

&

A

A n y q u e s t i o n ?
