판례데이터를 활용한
뉴스 댓글 고소 확률 예측

시계열자료분석팀

장다연 심현구 천예원 윤세인 이동기

목 차

1. 데이터 클렌징

1. 데이터 클렌징

추가 데이터 클렌징

1주차에서의 데이터 클렌징에 더해,
추가적인 클렌징이 필요하다고 판단

판례 댓글 클렌징
댓글과 발언 모두 사용
특수한 데이터에 대한 기준 제시 후 필터링

일반 댓글 클렌징
다양한 분야의 댓글 중 방송, 유머 분야의 댓글 선택

+

데이터 노이즈 제거
특정 시점 이전의 데이터 필터링

+

학회장 피드백 반영!
판례 댓글과 다른 성격의 데이터 제거

추가적인 데이터 클렌징 진행!

정승민… →

1. 데이터 클렌징

추가 데이터 클렌징

1주차에서의 데이터 클렌징에 더해,
추가적인 클렌징이 필요하다고 판단

판례 댓글 클렌징
댓글과 발언 모두 사용
특수한 데이터에 대한 기준 제시 후 필터링

+

데이터 노이즈 제거
특정 시점 이전의 데이터 필터링

일반 댓글 클렌징
다양한 분야의 댓글 중 방송, 유머 분야의 댓글 선택

+

판례 댓글과 다른 성격의 데이터 제거
학회장팀 피드백 반영

추가적인 데이터 클렌징 진행!

1. 데이터 클렌징

판례댓글 클렌징 | 노이즈 제거

사실 적시 및 허위 사실

사실 적시 혹은 **허위 사실**을
죄목으로 고소당한 경우,
모델이 진위여부를 판정할 수 없으므로
데이터셋에서 제거함.

피고 C은 2016. 7. 19. 15:48경 L 게시판에 원고를 "**남의 가정을 파탄에 이르게 한 상간녀**|'로 지칭하면서 "상간이란 도리에 어긋난 정을 나누는 행위, 단순히 배우자가 있는 사람과 연애만 하더라도 상간에 포함이 된다." "**남의 가정을 파탄에 이르게 한 상간녀는 오늘도 대로를 활보하고 위험한 인생의 짜릿함을 맛보기 위해 외줄 위로 올라간다. 상간녀 소송 좋은 방법이 핫 이슈가 되어 반갑다. 상간녀, ㅋㅋ 어감 한번 더럽네 ㅋㅋ 남의 돈으로 호텔가고 소고기 사먹고 좋았겠구나. 넌 최고니까!**"라는 내용의 글을 게시하여 허위 사실을 적시하여 원고의 명예를 훼손하였다.

문제가 되는 부분인 원고를 '상간녀'로 지칭한 부분은 허위사실이지만
모델이 이를 판단할 수 없으므로 제거!

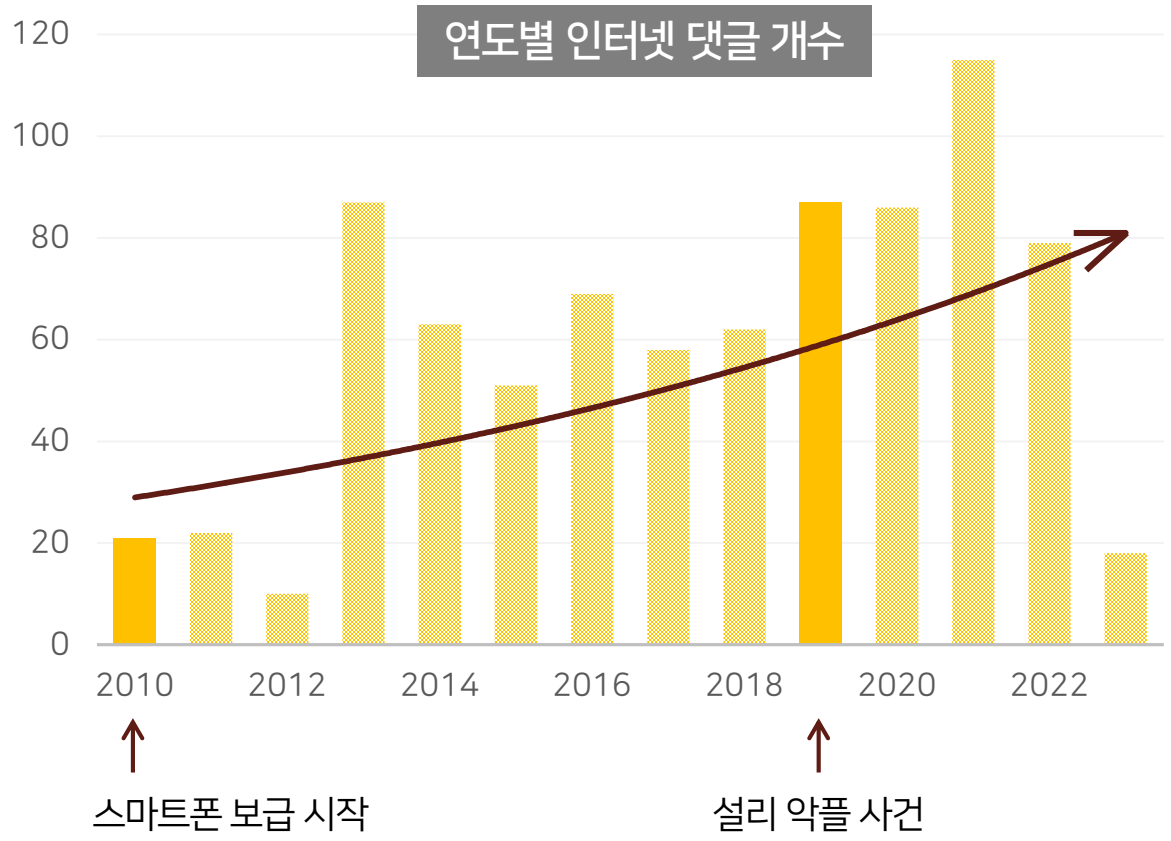
1. 데이터 클렌징

판례댓글 클렌징 | 과거 데이터

과거 데이터

'**실시간 댓글 데이터**의 고소 확률 예측' 이라는
목적에 맞게 **2010년 이전** 댓글 데이터는
분석 정확도를 위하여 데이터셋에서 제거함.

2008년 최진실 사건의 영향력이
이후 관련 재판에 반영되기까지의 2년의 시간을 고려해
2010년을 기준으로 채택



1. 데이터 클렌징

일반댓글 클렌징 | 이질적인 데이터 제거

판례 댓글과의 괴리가 큰 일반 댓글만을 대조군으로 사용할 경우,

일반댓글과 악플의 경계에 있는 모호한 댓글들의 예측력이 우려된다고 판단

야 이 씨X놈아 나가 죽어라

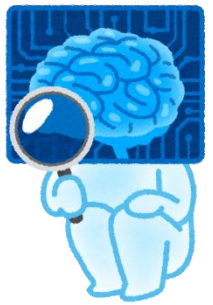
X새끼야

너무 재미있어요!

병X같은 XX 자살하시길

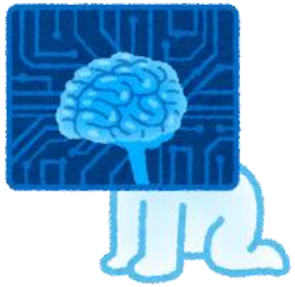
칼국수 맛있겠다!

사랑해요 엘지 ~



Label1과 Label0의 분류 기준은
이렇게 저렇게 하면 되겠다! 쉽네!

욕인지 칭찬인지 모르겠어요 …



니 항문에서 무지개빛이 나온다.

위와 같은 이유로 **모호한 댓글들에 대한 예측력 향상**을 위하여 이질적 데이터 필터링을 결정

1. 데이터 클렌징

일반댓글 클렌징 | 이질적인 데이터 제거

KR-SBERT의 STS를 이용하여
판례 댓글과 일반 댓글의
1:多 유사도를 계산

↓

판례 댓글별로
유사도 상위 5개의 평균을 구한 후
그 값이 높은 상위 855개만을 추출

판례 댓글 – 일반댓글 多:1 유사도 계산 결과

	일반댓글	판례댓글		score
1	,발...	1	아이 씨발	0.5253
		...	...	...
		5	씨바	0.5176
2	04년에 무서워서 제대로...	1	보고나서 내킬 때 수위 높은 ...	0.5110
		...	...	
		5	위에 보셨듯이 이번 고소사건...	0.4474
3	1 2랑 3은 좀 많이 다른...	1	감독이 작품명이 'G'인데 'L'이라...	0.4587
		...	...	
		5	김 대통령이 당선되면 연기자가...	0.3690
...	...		...	...

유사도 상위 5개 댓글의 스코어

1. 데이터 클렌징

일반댓글 클렌징 | 이질적인 데이터 제거

최종적으로 실제 고소를 당한 ' 판례 댓글 ' 과

유사성을 띄는 댓글을 대조군으로 사용함으로써

모델이 모호한 댓글에 대해서도 더 정확하게 판단할 수 있도록 함!

1. 데이터 클렌징

일반댓글 클렌징 | 학회장팀 피드백

현실과의 괴리가 너무 큰 거 아닌가요?

승민

"판례 댓글과 일반 댓글을 1:1로 설정할 경우
고소 당할 확률을 50%라고 가정하는 것이 아닌지?"

전체 댓글 수 대비 실제 댓글이 고소당하는 비율은 0에 수렴함.
즉, **고소 확률 자체를 데이터셋 구성에 반영하는 것은 불가능**!

따라서 **균형 있는 모델 학습**을 궁극적 목표로 삼아
판례 댓글과 일반 댓글의 비율을 1:1로 설정하여 데이터셋 구성

다연

오히려 고소 확률을 적극적으로 잡아냄으로써
악플러들에게 경각심을 주기 위한 연구의 목적을 달성할 수 있음!

1. 데이터 클렌징

최종 데이터셋

추가적인 데이터 클렌징 후 최종 데이터셋

	text	comments	고소 label
1	고등군사법원 2016.6.30 …	네 여자친구랑 섹스해도 되냐?	1
2	고등군사법원 2017.1.25 …	이게 청소한 거냐, 이새끼들이 미쳤네, 초임하사 새끼들이 벌써부터 풀려가지고, 니…	1
3	고등군사법원 2017. 1. 25 …	넌 새끼야 왜 똑같이 가르쳤는데 쟤보다 못 하냐, 너 하사 아니가, 하사가 병들을 …	1
…	…	…	…
1708	NaN	아 민석이형 길거리에서 봤는데 부끄러워서 …	0
1709	NaN	걍 이건 괴롭히기잖아	0
1710	NaN	얼굴 가리고 보면 개멋있음	0

1행~855행 : 판례 댓글
856행~1710행 : 일반 댓글

2. 파생변수

2. 파생변수

파생변수 생성

임베딩된 댓글 데이터 입력만으로

모델이 학습할 수 없는 정보들을 파생변수로 추가

파생변수를 추가하여

모델에 반영되지 못해 error가 된 정보량을

모델 변수로 편입시킴으로서 예측력 향상을 기대할 수 있음

방긋 …

- 댓글 감성분석
- 댓글 작성 년도
- 욕설 포함 확률
- 댓글 길이

2. 파생변수

파생변수 생성

파생변수 vs 임베딩 벡터

임베딩된 댓글 데이터 입력만으로

모델이 학습할 수 없는 정보들을 파생변수로 추가

텍스트에서 나온 결과들로 파생변수를 구성하기 때문에,

이미 임베딩 벡터가 이에 대한 정보를 반영하고 있을수 있을 수 있음

모델 변수로 편입시킴으로서 예측력 향상을 기대할 수 있음

방긋 …

파생변수의 중요도가 임베딩 벡터의 그것보다 낮을 경우,

해당 파생변수는 사용하지 않기로 결정

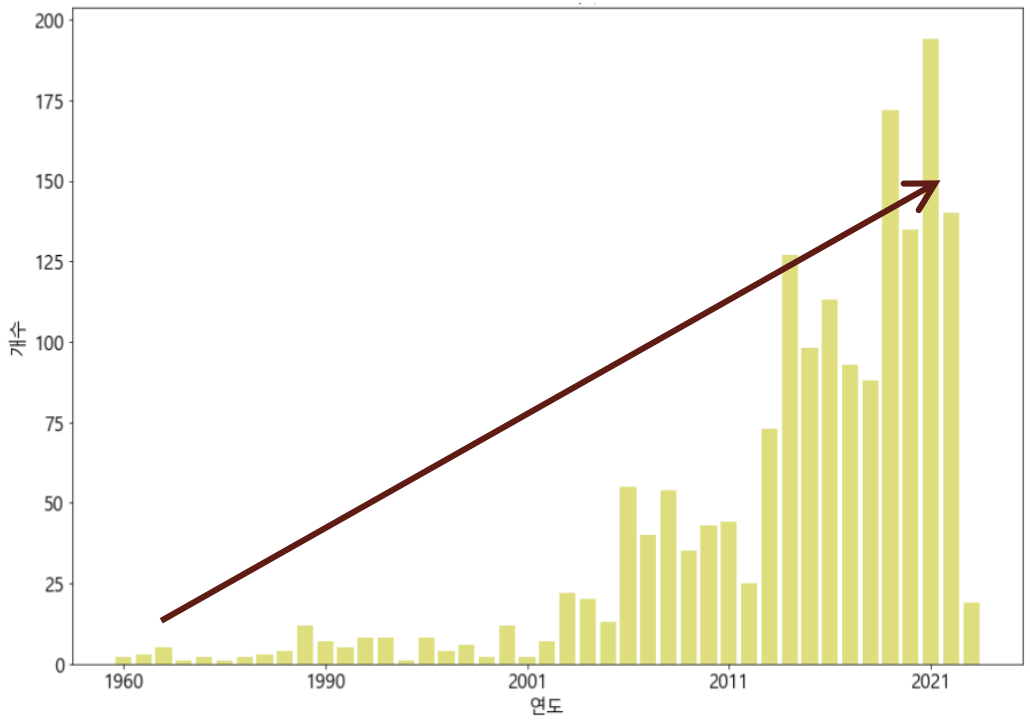
댓글 내 욕설 여부

댓글 길이

2. 파생변수

파생변수 | 작성 년도

연도별 판례 댓글 개수



시간이 지남에 따라 **판례 댓글의 개수가 증가**하고 있음을 확인

["댓글에 `무뇌아` 단어 썼으면 모욕죄"]

김 씨는 지난 2013년 한 인터넷 카페에 게시된 글에 윤모씨를 **무뇌아로 지칭하는 댓글을 달아 윤씨를 모욕한 혐의**를 받고 있다.

(중략)

"같은 단어라도 댓글 같은 짧은 글에서는 전체 맥락을 살피기 어려워 상대방을 비난하는 단어를 쓰면 모욕죄가 될 개연성이 크다"

최신판례는 **댓글로 모욕한 경우의 범죄성립요건을 발언의 경우보다 완화**하는 경향을 보임

2. 파생변수

파생변수 | 작성 년도

악플과 판례의 경향성을 반영하여

판결연도와 작성연도를 변수로 추가!

판례 댓글	일반 댓글
판례에 명시된 판결 년도	댓글 작성 년도

가설 : **최근 댓글일수록 고소 확률이 올라갈 것이다.**

더불어, 추가된 파생변수에 대해 가설검정하여

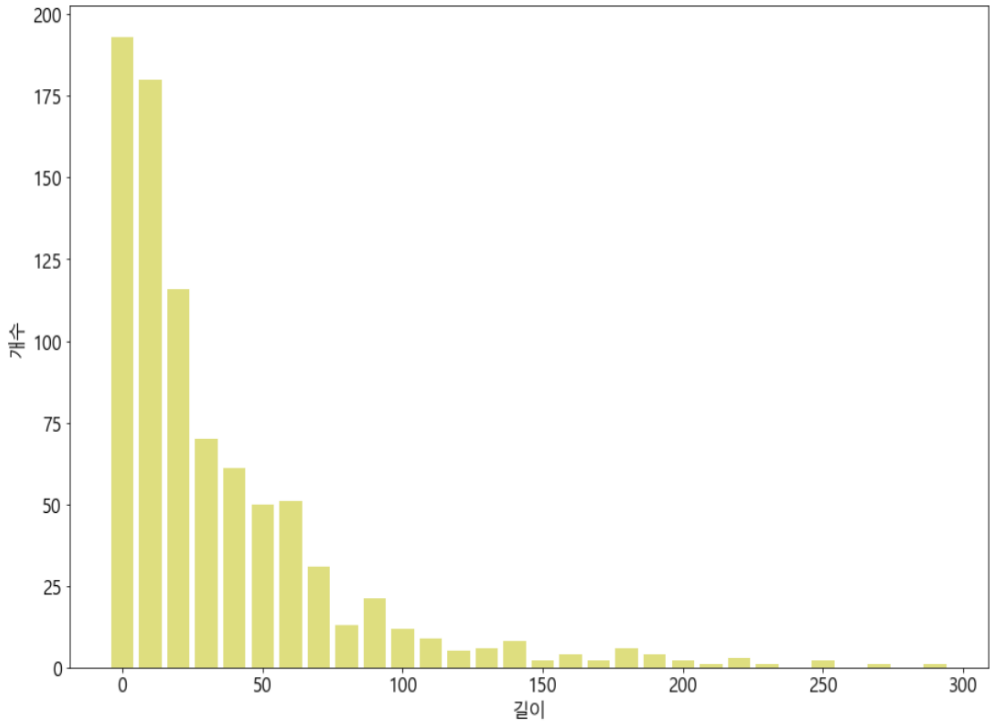
변수 유의성 확인 예정

과연…

2. 파생변수

파생변수 | 댓글 길이

길이별 댓글 개수



댓글 길이가 짧으면 **특정성**, **공연성**과 같은

범죄구성요소를 모두 반영하기 어렵지 않을까?!

가설 : **댓글 길이가 길수록 고소 확률이 올라갈 것이다.**

댓글 길이가 고소 확률에 유의미한 영향을 미칠 것

이라는 가정 하에 댓글의 길이를 변수로 추가함

2. 파생변수

파생변수 | 댓글의 성격을 반영한 변수

피셋 도겸

일반 댓글 데이터셋 중 **악플의 성격을 띄는 경우**는
어떻게 구분하실 건가요 … ?

댓글이 가진 주관적인 속성을 학습에 반영할 수 있도록
LDA, KoBERT **감성분석**을 통한 파생변수 생성을 고려!

후 …

피셋 윤아

2. 파생변수

파생변수 | 토픽모델링

토픽 모델링

문장들의 Corpus에 내재되어 있는 토픽을 끌어내는 데 쓰이는 방법

전체 문서를 하나의 주제로 보고 **주제를 구성하는 토픽**을 찾아내 문장을 분류

토픽 모델링 모델

LDA

LSA

BERT

유사한 토픽을 가진 문장을 하나로 묶어 **다중 분류 시도**

2. 파생변수

파생변수 | 토픽모델링

토픽 모델링

문장들의 Corpus에 내재되어 있는 토픽을 끌어내는 데 쓰이는 방법

전체 문서를 하나의 주제로 보고 **주제를 구성하는 토픽**을 찾아내 문장을 분류

피셋 도겸

LDA는 단어 중심 모델이라 악플 관련 단어들이
하나의 토픽으로 잡혀야 하는데 … 쉽진 않아보이네요 …

KoBERT 감성분석 해볼게요 …

주륵 …

피셋 하와수

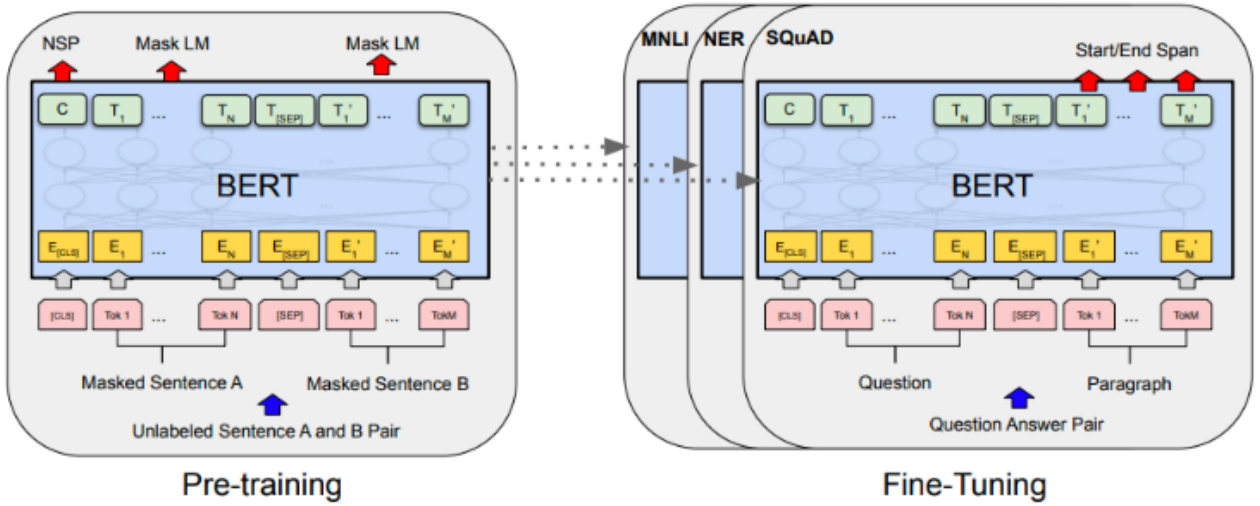
2. 파생변수

파생변수 | KoBERT 감성분석

BERT

사전 학습된 대용량의 레이블링 되지 않는 데이터를 이용하여 언어모델을 학습하는 방법

Transformer 구조에 encoder를 여러 층 더한 구조



Bert 모델의 장점

1. 문맥을 반영한 임베딩 (Contenxtual Embedding) 사용
2. 서브워드 토크나이저로 자주 사용되는 단어와 그렇지 않은 단어를 다른 방식으로 토큰화

감성분석에 좋은 성능을 보임!

2. 파생변수

파생변수 | KoBERT 감성분석

KoBERT

SKBrain에서 개발한 모델

BERT 모델에서 **한국어 데이터를 추가적**으로 학습시킴

→ 한국어 위키의 5백만개의 문장 + 54만개의 단어 사전학습

흐뭇…

KoBERT를 이용해 댓글의 감정을 변수로 추가하자!

가설 : **부정적인 감정**으로 분석된 댓글일수록
고소 확률이 올라갈 것이다.

2. 파생변수

파생변수 | KoBERT 감성분석

학습 데이터셋

AIhub에서 제공하는 한국어 감성 대화 말뭉치

기쁨, 슬픔, 분노, 불안, 당황, 상처의 6가지 감정으로 라벨링이 완료된 데이터

학습 문장	감정
요즘 직장생활이 너무 편하고 좋은 것 같아!	기쁨
오늘 회사에서 큰 실수를 한 것 같아.	슬픔
일은 왜 해도 해도 끝이 없을까? 화가 난다.	분노
큰일이야! 중요한 물건이 사라졌어.	불안
길을 가다가 만난 아주머니가 취업에 대해 물어보셨어	당황
결혼은 현실이니까. 돈이 없어.	상처

2. 파생변수

파생변수 | KoBERT 감성분석

학습 데이터셋

AIhub에서 제공하는 한국어 감성 대화 말뭉치

기쁨, 슬픔, 분노, 불안, 당황, 상처의 6가지 감정으로 라벨링이 완료된 데이터

학습 문장	감정
요즘 직장생활이 너무 편하고 좋은 것 같아!	기쁨
오늘 회사에서 큰 실수를 한 것 같아.	슬픔
일은 왜 해도 해도 [illegible]	분노
큰일이야! 중요한 물건이 사라졌어.	불안
길을 가다가 만난 아주머니가 취업에 대해 물어보셨어	당황
결혼은 현실이니까. 돈이 없어.	상처

Fine tuning 후 예측 진행!

2. 파생변수

파생변수 | KoBERT 감성분석

댓글	감정
아 기분 좋아져 영상이	기쁨
아니근데 사람 마음으로 장난치지 말지	슬픔
더러운 새끼 지랄하네. 개새끼 쓰레기 새끼	분노
전쟁이라고 위기감 조성시키고..	불안
어 내가 아는 현석이형이 없는데	당황
그렇구나 알겠어	상처

기쁨	슬픔	분노	불안	당황	상처
1	0	0	0	0	0
0	1	0	0	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1

One-hot encoding을 통해 변수 생성

2. 파생변수

파생변수 | 욕설 포함 확률

판례 키워드
정보통신망이용촉진및정보보호등에 관한법률위반(명예훼손), 모욕죄, 비방, 댓글, 통신매체이용음란죄

판례 댓글
그 좇같은 새끼, 개같은 새끼, 쌍놈의 새끼라고
아 십ㄹㄹ발ㄹㄹㄹㄹ 좇같아요
돼지새끼, 허리병신새끼, 뇌까지 장애 있어?
씨발, 나이 똥구멍으로 쳐먹지마
이 미친놈아 너 나 지금 엿먹일려고 작정했냐
나만 차별하는 건가 지랄한다 짜증나게

모욕, 비방 등의 키워드로 댓글을 추출해 **욕설이 포함된 댓글**이 다수 존재

MFCC 임베딩 기반 모델로 댓글 내의 욕설 포함 확률을 탐지하여

파생변수로 활용하자!

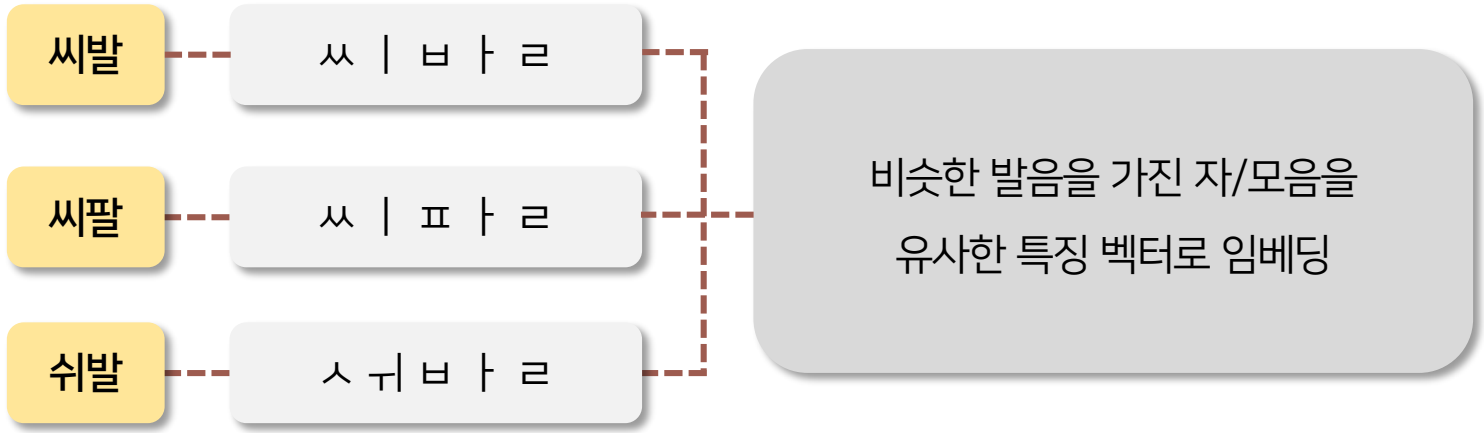
2. 파생변수

파생변수 | 욕설 포함 확률

MFCC(Mel-Frequency Cepstral Coefficient)

음성 데이터를 특징 벡터화하는 알고리즘

각 주파수마다 다른 weight를 가진 필터를 통해 음고를 계산하는 방식



MFCC 알고리즘을 응용한 댓글 임베딩으로 다양한 **욕설 파생형**을 탐지할 수 있도록 함

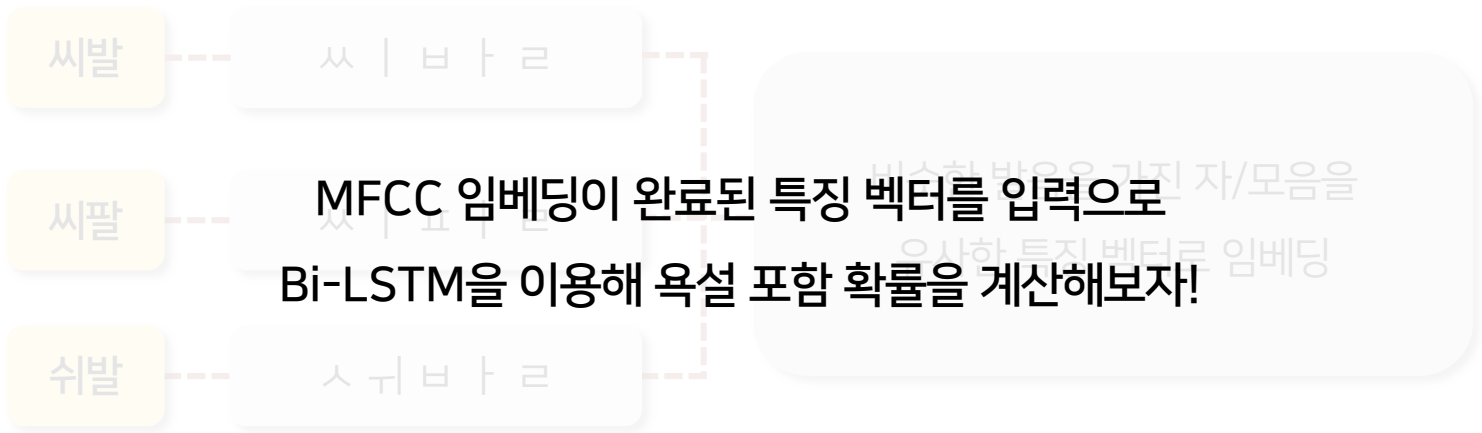
2. 파생변수

파생변수 | 욕설 포함 확률

MFCC(Mel-Frequency Cepstral Coefficient)

음성 데이터를 특징 벡터화하는 알고리즘

각 주파수마다 다른 weight를 가진 필터를 통해 음고를 계산하는 방식



MFCC 임베딩이 완료된 특징 벡터를 입력으로

Bi-LSTM을 이용해 욕설 포함 확률을 계산해보자!

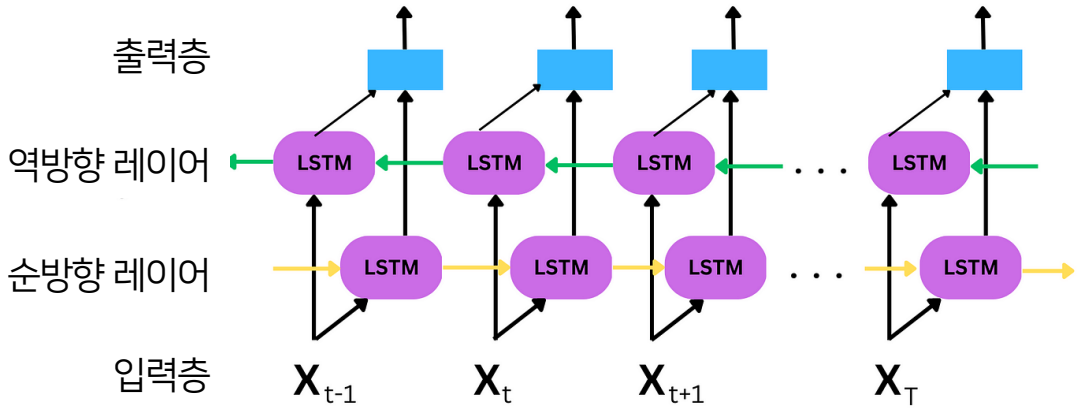
MFCC 알고리즘을 응용한 댓글 임베딩으로 다양한 **욕설 파생형**을 탐지할 수 있도록 함

2. 파생변수

파생변수 | 욕설 포함 확률

Bi-LSTM (Bidirectional LSTM)

정방향 (forward) LSTM과 역방향 (backward) LSTM 계층을 포함한 모델로,
문장 시퀀스의 이전 부분과 이후 부분을 모두 고려하여 학습한다는 점에서
자연어 처리 모델로서 효과적이다.



MFCC 임베딩 + Bi-LSTM 기반 Classifier를 이용하여
욕설 포함 확률 추가변수 생성 완료 ~

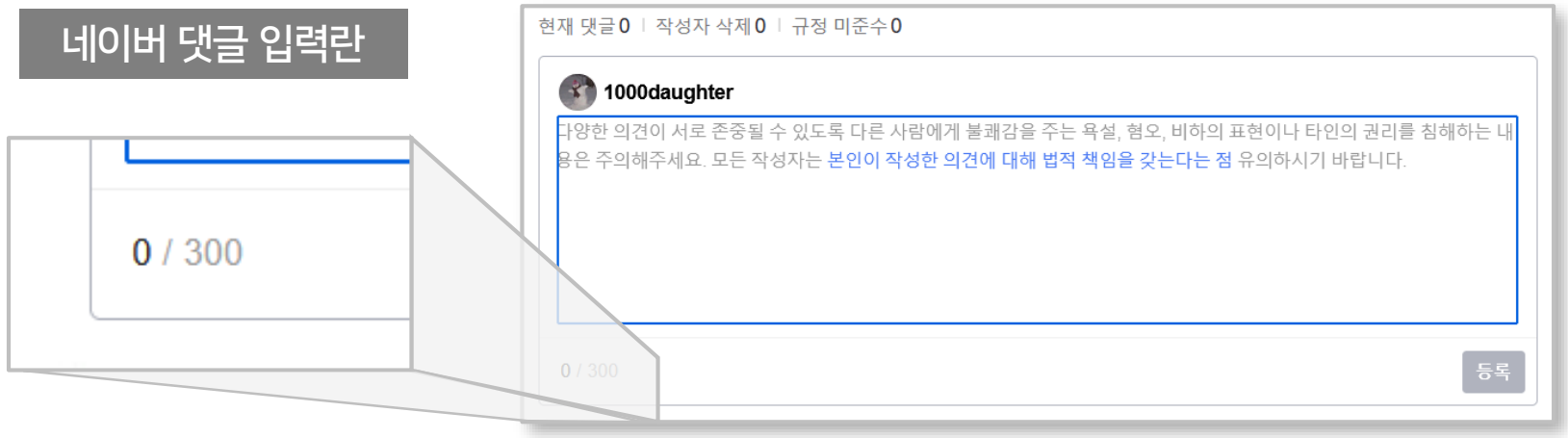
3. 텍스트 전처리

3. 텍스트 전처리

텍스트 전처리 | 글자수 제한

우리아빠가 지어준 아이디임 시비 ㄴㄴ

네이버 댓글 입력란



길이가 너무 긴 댓글은 **댓글의 특성을 잃었다고 판단**

네이버 뉴스 기사 댓글의 **300자 제한** 기준을 따라 300자를 초과하는 댓글 데이터 삭제!

3. 텍스트 전처리

텍스트 전처리 | 반복 음절 축약

댓글이 **긴 의성어**를 포함하는 경우 **BERT 사용시 윈도우에 악영향**

피고인은 2019. 6. 4. 19:06경 인터넷 포털사이트 C 카페에 닉네임 'B'으로 접속하여 피해자 D(24세)이 평소 그 카페에서 사용하는 자신의 휴대전화번호와 함께 게재한 신발 판매글에 대하여 **"ㅋㅋㅋㅋ 짭을 찐처럼 쳐팔고 앉아잇네 사기꾼새끼 ㅉㅉㅉㅉ"**라는 댓글을 게재하여 공연히 그를 모욕하였다.

여러 번 반복되는 음절을 **최대 2회 반복**되도록 한정하여 축약시킴

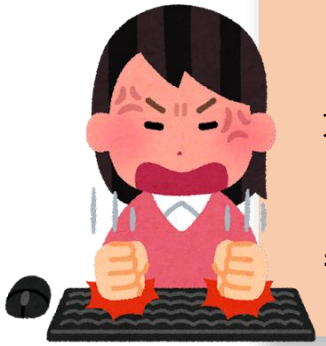
ㅋㅋ 짭을 찐처럼 쳐팔고 앉아있네 사기꾼새끼 ㅉㅉ

3. 텍스트 전처리

텍스트 전처리 | 영어, 숫자, 특수문자 제거

모델 학습 과정에서 **데이터의 일관성**을 유지하고,

한국어 댓글에 보다 민감하게 반응할 수 있도록 **한국어만을 추출**하여 사용



피고인은 피해자 C과 인터넷 블로그에서 알게 되어 수차례 의견 대립으로 다투게 되자, 2014. 11. 19. 13:18경 자신의 핸드폰을 이용하여 피해자가 사용하는 핸드폰으로 "**그 아스퍼거가 너라는거 인정하고 말고는 여기서 아무 factor가 아니라는 것을 알텐데 아스퍼거야ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ**"라는 메시지를 전송한 것을 비롯하여 … 피해자의 명예를 훼손하였다.

‘factor’라는 영단어가 모델 학습 과정에 악영향을 줄 수 있으므로 생략

그 아스퍼거가 너라는거 인정하고 말고는 여기서 아무가 아니라는 … (후략)

4. 변수 선택 및 모델 선정

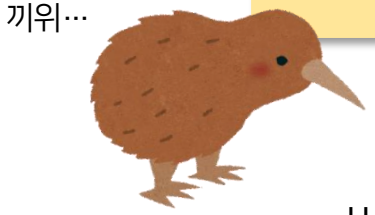
4. 변수 선택 및 모델 선정

토큰화 및 임베딩

토크나이저

KIWI

OKT



댓글이라는 데이터의 특성을 반영하여

비표준어, 속어 등의 처리에 강한 토크나이저를 채택!

임베딩 모델

Word2Vec

FastText

KLUE-BERT

단어 기반 임베딩과 문장 기반 임베딩 모델의 성능을 비교

4. 변수 선택 및 모델 선정

토큰화 및 임베딩

토크나이저

KIWI

OKT

끼위…

댓글이라는 데이터의 특성을 반영하여

비표준어, 속어 등의 처리에 강한 토크나이저를

마 주분에 집중 안하나!

임베딩 모델

Word2Vec

FastText

KLUE-BERT

잠깐! KLUE-BERT에 대해 빠르게 알고 넘어가자!

단어 기반 임베딩과 문장 기반 임베딩 모델의 성능을 비교

4. 변수 선택 및 모델 선정

토큰화 및 임베딩

KLUE-BERT

한국어로 **사전 훈련**된 BERT 모델

주제 분류, 의미론적 텍스트 유사성,

자연어 추론, KLUE 벤치마크 등에 활용 가능

사전학습 한국어 말뭉치
국립국어원의 '모두코퍼스 '
대규모 다국어 웹 크롤링 말뭉치 CC-100_Kor
나무위키
NEWSCRAWL
청와대 국민청원

KLUE-BERT 토크나이저의 encode 메서드를 이용하여

별도의 과정 없이 토큰화와 임베딩을 한번에 가능!

4. 변수 선택 및 모델 선정

토큰화 및 임베딩

Okt 토크나이징 결과

Okt 토큰화 전	씨발, 나이 똥구멍으로 쳐먹지마
Okt 토큰화 후	“씨발”, “,”, “나이“, “똥구멍“, “쳐”, “먹지마”

Word2Vec 임베딩	LGBM	RandomForest	Logistic
	0.74	0.77	0.73

FastText 임베딩	LGBM	RandomForest	Logistic
	0.73	0.77	0.74

4. 변수 선택 및 모델 선정

토큰화 및 임베딩

Kiwi 토크나이징 결과

Kiwi 토큰화 전	“씨발, 나이 똥구멍으로 쳐먹지마”
Kiwi 토큰화 후	'씨발', '나이', '똥구멍', '치', '먹', '말'

Word2Vec 임베딩	LGBM	RandomForest	Logistic
	0.79	0.77	0.58

FastText 임베딩	LGBM	RandomForest	Logistic
	0.81	0.78	0.58

4. 변수 선택 및 모델 선정

토큰화 및 임베딩

KLUE-BERT 토크나이징 + 임베딩 결과

KLUE-BERT 토큰화 및 임베딩 전	"씨발, 나이 똥구멍으로 쳐먹지마"

KLUE-BERT 임베딩	LGBM	RandomForest	Logistic
	0.91	0.90	0.88

4. 변수 선택 및 모델 선정

모델링 조합 선택

다양한 토큰화, 임베딩, 예측 모델조합으로
성능 비교 후 최적의 조합을 선택!

노트북아 죽지마 …

토크나이저	임베딩 모델	예측 모델	
	Word2Vec	Logistic Regression	0.88
Kiwi	FastText	LGBM	**0.93**
Okt	KLUE-BERT	DNN	0.87
		CNN	0.77
		RandomForest	0.90

4. 변수 선택 및 모델 선정

모델링 조합 선택

다양한 토큰화, 임베딩, 예측 모델조합으로
성능 비교 후 최적의 조합을 선택!

노트북아 죽지마 …

F1 score가 가장 높았던

KLUE-BERT 임베딩 + LightGBMClassifier 분류 모델을

예측에 사용하기로 최종 결정!

Logistic Regression

LGBM

DNN

CNN

Word2Vec

FastText

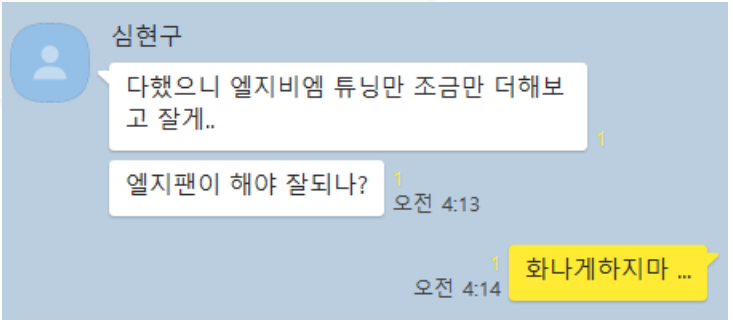
KLUE-BERT

Kiwi

Okt

토크나이저

임베딩 모델



4. 변수 선택 및 모델 선정

변수 선택

임베딩된 댓글 데이터 입력만으로

모델이 학습할 수 없는 정보들을 파생변수로 추가

- 댓글 감성분석
- 댓글 작성 년도
- 욕설 포함 확률
- 댓글 길이

XAI를 통해 최종 변수를 선택해보자!

4. 변수 선택 및 모델 선정

XAI

설명가능한 인공지능(eXplainable AI)

머신러닝 / 딥러닝 알고리즘으로 생성된 결과와 출력물을

이해하고 **신뢰**할 수 있게 만드는 방법

XAI 알고리즘

- 개별 변수 중요도 확인
  - Feature Importances
  - SHAP
- 시각화
  - LIME

4. 변수 선택 및 모델 선정

Feature Importances

Feature Importances

트리 기반 모델에서 각 특징이 모델 예측에 얼마나 기여하는지를 나타내는 지표

트리 노드에서의 분기 중요도와 **트리에서의 불순도 감소 합산**을 이용해

변수의 상대적인 중요도를 측정

통데마 수업 中

Feature Importance를 이용한 변수선택은 위험하니까

개별 변수의 중요도를 대략적으로 확인하는 지표로서 활용하세요~

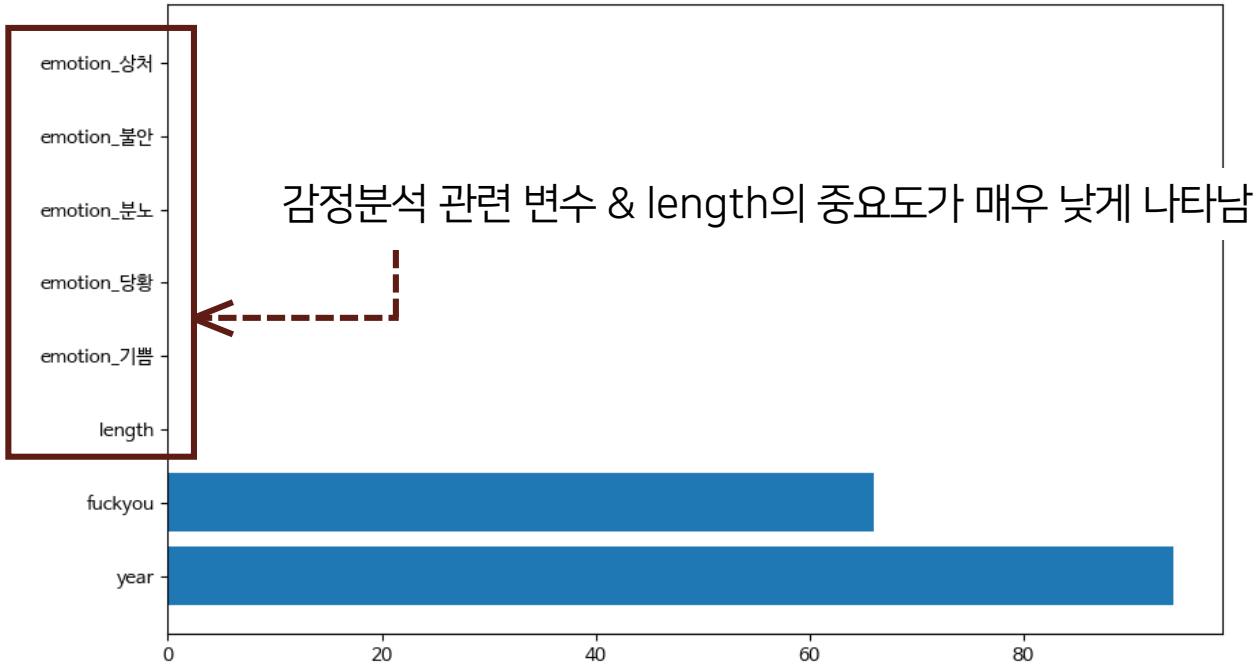
(여태까지 그렇게 했는데 …) 네~

피셋 하와수

4. 변수 선택 및 모델 선정

Feature Importances

LGBM기반 파생변수 Feature Importances



감정분석 관련 변수 & length의 중요도가 매우 낮게 나타남

중요도가 낮게 나타난 파생변수는 **이미 임베딩 벡터에 반영**되어 있다고 판단할 수 있음.

SHAP을 통해 확인한 변수중요도와 비슷한지 확인 후 제거할 예정

4. 변수 선택 및 모델 선정

SHAP

SHAP

Shapley value라는 게임 이론을 바탕으로 하며,

각 특징에 대한 Shapley 값은 그 **특징이 모델 예측에 기여하는 정도**를 나타냄

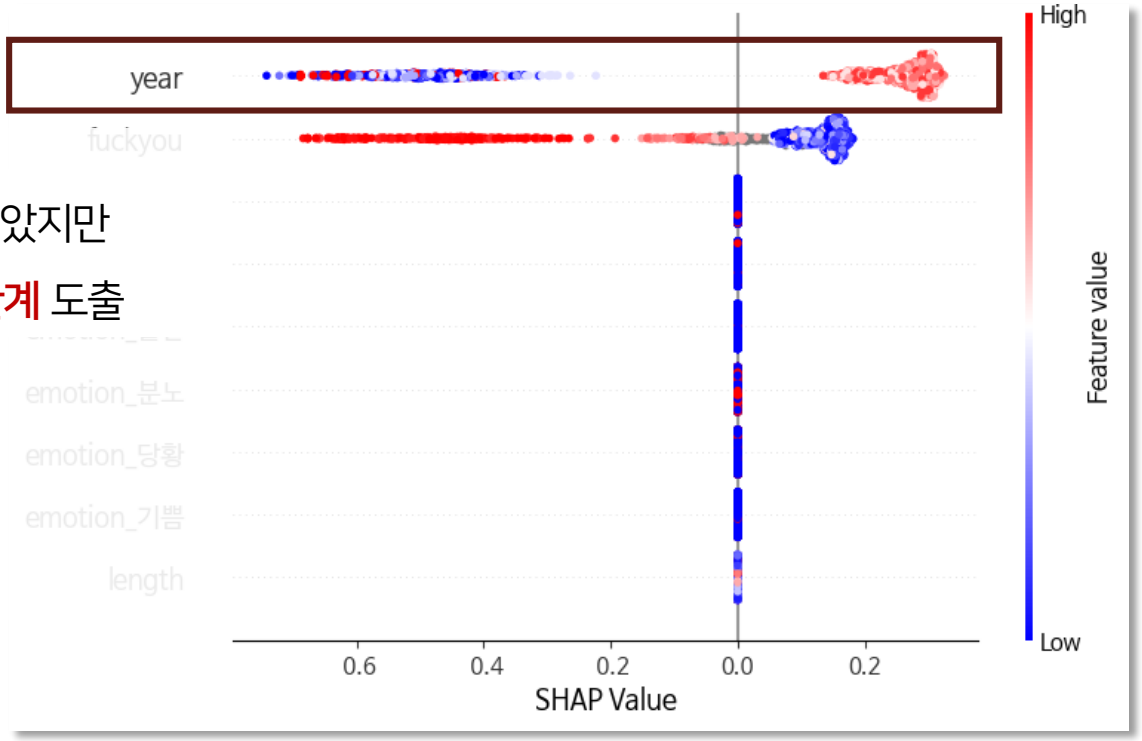
전체적인 기여도와 Fairness를 고려해 **모델의 예측에 어떤 특징이 어떻게 기여하는지** 설명할 수 있음

- SHAP의 출력 결과 중 변수중요도를 활용해 **변수선택**을 진행
- 변수가 예측에 영향을 미치는 **방향성** 확인

4. 변수 선택 및 모델 선정

변수선택 | SHAP | year

SHAP Dot Plot



'year'는 변수중요도는 높았지만

예상과 반대로 **음의 상관관계** 도출

판례댓글은 2010년부터, 일반댓글은 2016년부터로 구성되어 있기 때문이라고 추측

→ 판례댓글을 2016년부터로 재구성 후 모델검정(LRT) 진행

4. 변수 선택 및 모델 선정

가설검정

가설검정1 – 모델 유의성

H_0: full model과 reduced model에 유의미한 차이가 없다

H_1: full model과 reduced model에 유의미한 차이가 있다

가설검정2 – 변수 유의성

H_0: 해당 변수에 따른 유의미한 차이는 존재하지 않는다

H_1:해당 변수에 따른 유의미한 차이가 존재한다

4. 변수 선택 및 모델 선정

가설검정

가설검정1 – 모델 유의성

H_0: full model과 reduced model에 유의미한 차이가 없다

H_1: full model과 reduced model에 유의미한 차이가 있다

가설검정2 – 변수 유의성

R의 lrtest()를 통해 모델 유의성을 검정해보자!

lrtest(full_model, reduced_model)

H_0: 해당 변수에 따른 유의미한 차이는 존재하지 않는다

H_1:해당 변수에 따른 유의미한 차이가 존재한다

4. 변수 선택 및 모델 선정

가설검정

LRT (Likelihood Ratio Test)

모델의 가능도 비교를 통해
가설을 검정하는 방법

lrtest(full_model,reduced_model)

검정통계량: $L = -2\log\left(\frac{L(\hat{\beta})}{L(\hat{\beta}^{MLE})}\right) \sim X^2_{df}$

$L(\hat{\beta})$: 귀무가설 하에서의 가능도 함수 최대값

$L(\hat{\beta}^{MLE})$: 실제 MLE

- 두 모델의 **가능도 차이가 작다면** 추정값이 MLE에 가까워져 **귀무가설 기각할 수 없음**
- 두 모델의 **가능도 차이가 크다면** 검정통계량이 커져 **귀무가설 기각**

4. 변수 선택 및 모델 선정

가설검정

H0:Reduced Model 채택 H1: Full model 채택					
#DF	LogLik	DF	Chisq	Pr(>Chisq)	
769	-8.9724				
770	-7.4675	1	3.0098	0.08276	.
Signif. codes : 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1					

- Full model: 임베딩 벡터 + year
- Reduced model: 임베딩 벡터

year에 대한 모델적합성검정(Likelihood Ratio Test)을 한 결과 p-value = 0.0827

0.1 유의수준 하에서 귀무가설을 기각해 full model 채택!

추가로 year 변수에 대한 유의성 검정을 진행함

4. 변수 선택 및 모델 선정

가설검정

가설검정1 – 모델 유의성

H_0: full model과 reduced model에 유의미가 차이가 없다

H_1: full model과 reduced model에 유의미가 차이가 있다

Logistic Regression **계수**를 통해 변수 유의성 검정!

가설검정2 – 변수 유의성

H_0: 해당 변수에 따른 유의미한 차이는 존재하지 않는다

H_1:해당 변수에 따른 유의미한 차이가 존재한다

4. 변수 선택 및 모델 선정

가설검정

H0: year의 회귀계수가 유의하지 않다 H1: year의 회귀계수가 유의하다					
	Estimate	Std.Error	Z value	Pr(>|z|)	
year	1.20	0.796	1.51	0.1316	

변수 자체에 대한 t-test 결과 p-value = 0.1316

유의수준으로 설정한 0.1 경계에 존재함을 확인

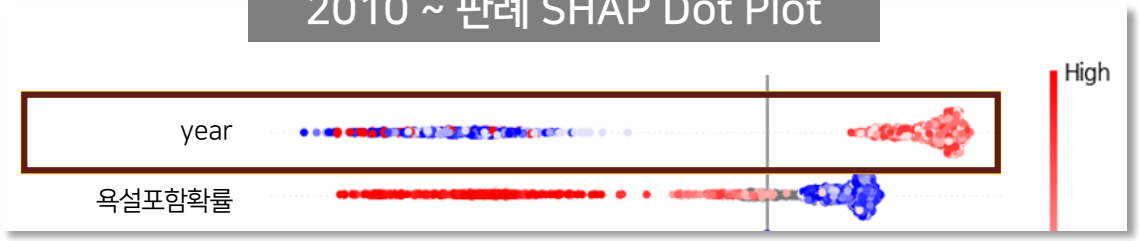
섣불리 귀무가설을 채택/기각할 수 없다고 판단해

다시 한 번 XAI를 참고해서 변수선택하기로 함!

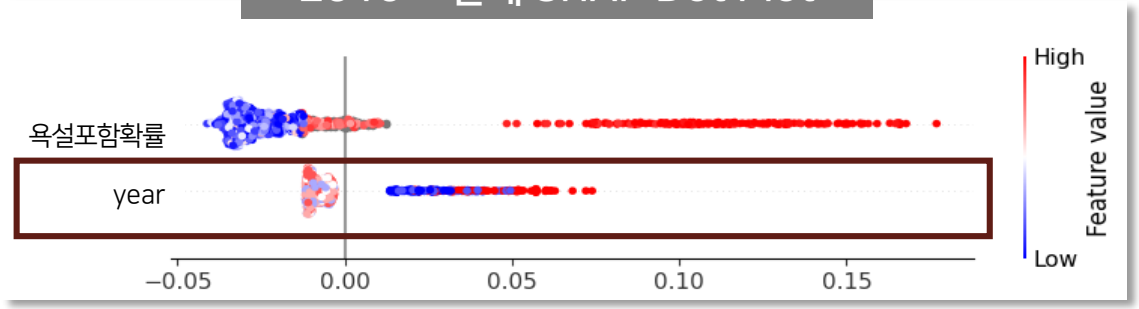
4. 변수 선택 및 모델 선정

변수선택 | SHAP | year

2010 ~ 판례 SHAP Dot Plot



2016 ~ 판례 SHAP Dot Plot



판례댓글을 2016년부터로 재구성한 후 성능 F1 score = 0.9059

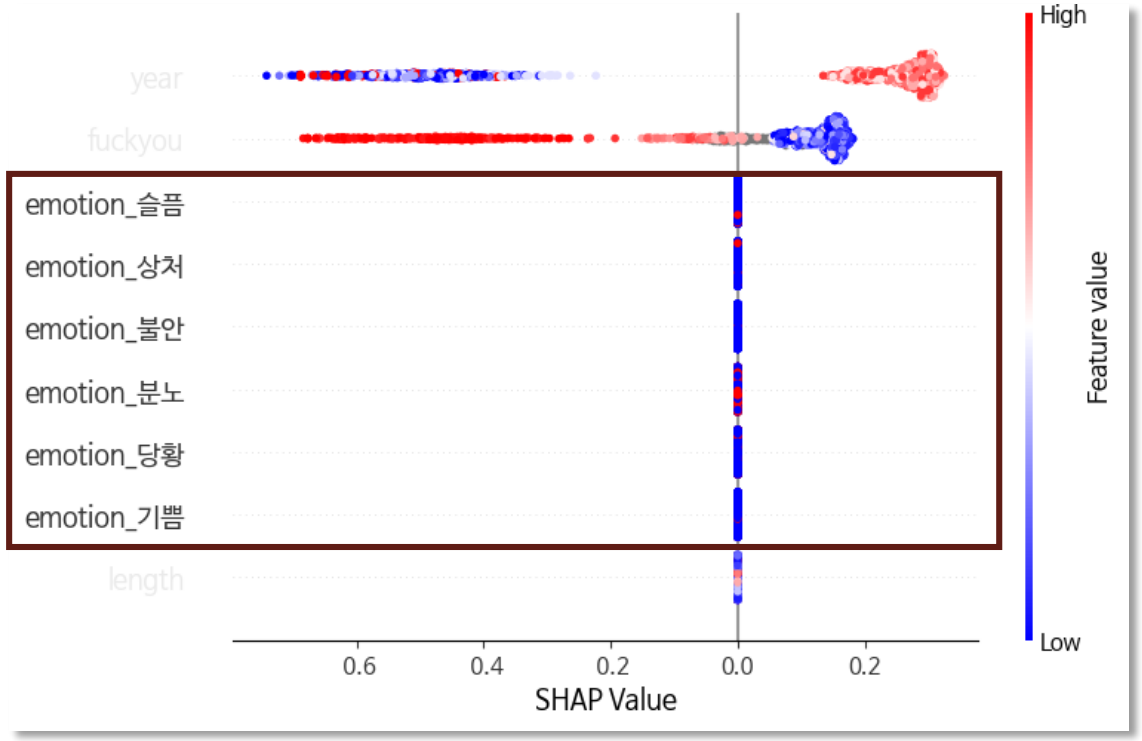
SHAP 결과를 확인했을 때 고소여부와 **양의 상관관계**로 방향이 바뀜!

'year'를 최종 파생변수로 채택!

4. 변수 선택 및 모델 선정

변수선택 | SHAP | Emotion

SHAP Dot Plot



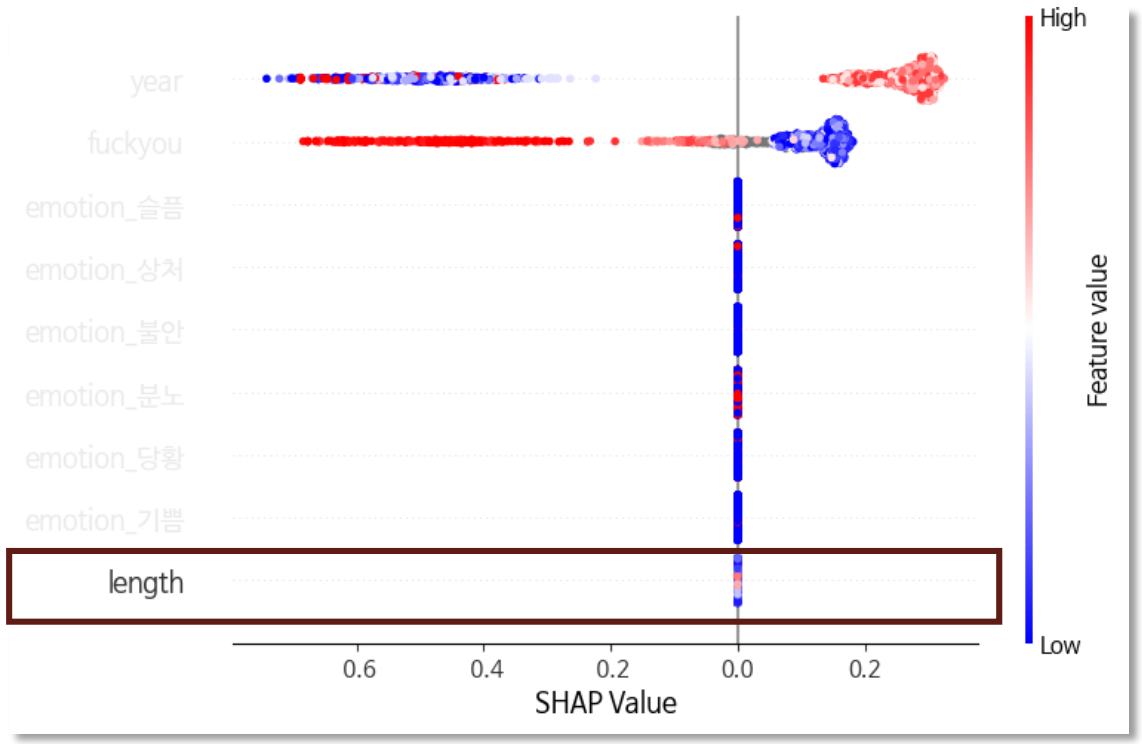
Emotion 파생벡터 : feature importances와 같이 낮은 변수중요도

→ 임베딩벡터에 포함된 정보로 해석, 제거

4. 변수 선택 및 모델 선정

변수선택 | SHAP | 댓글 길이

SHAP Dot Plot



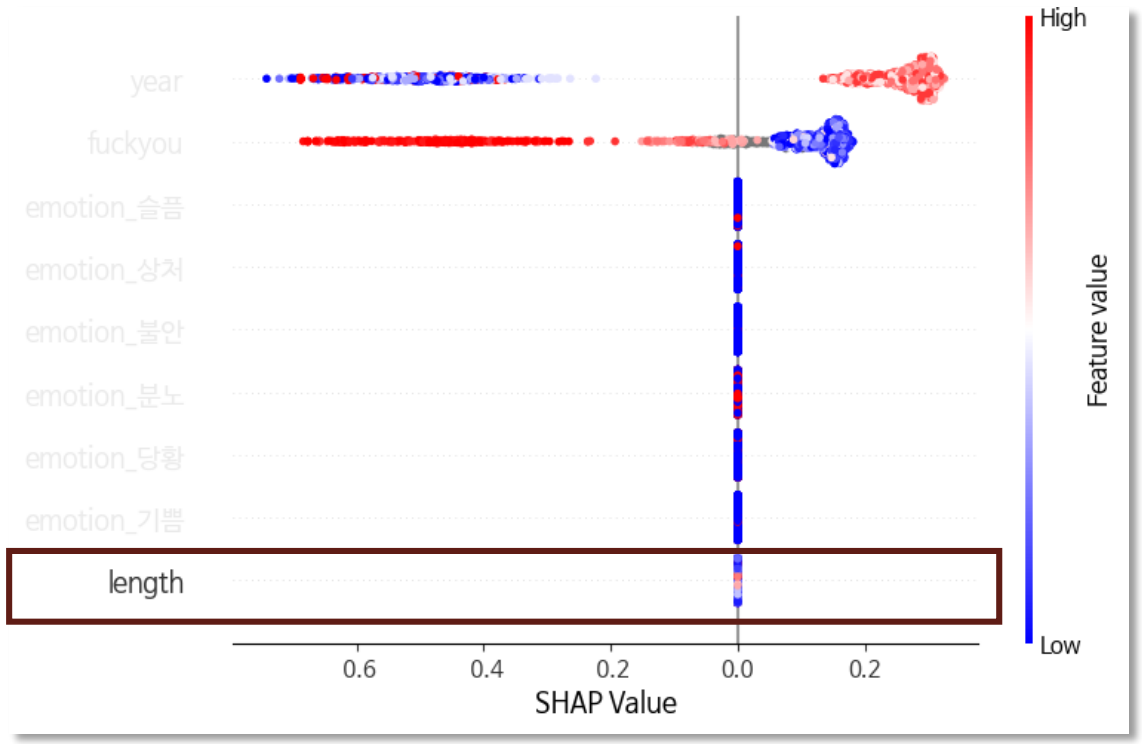
댓글길이는 너무 짧은 문장에는 범죄구성요건이 다 포함되기 어렵다고 생각해서 추가한 파생변수였는데,

Feature Importances와 마찬가지로 **중요도도 매우 낮고** 추가 시 **오히려 모델 성능이 하락**

4. 변수 선택 및 모델 선정

변수선택 | SHAP | 댓글 길이

SHAP Dot Plot



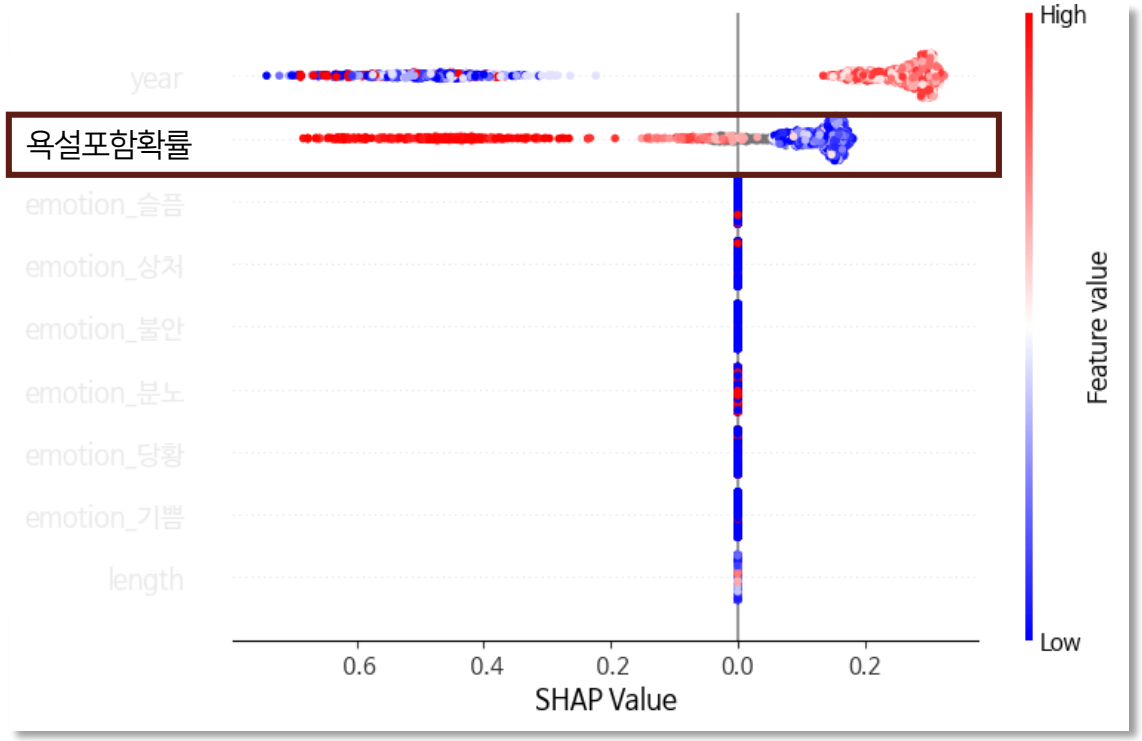
판례댓글데이터 중에도 "국민호텔녀", "씨발련아" 등 짧은 댓글이 꽤 관측되기 때문이라고 해석함.

최종적으로 **댓글 길이 파생변수를 학습 방해 변수**로 판단하여 삭제함.

4. 변수 선택 및 모델 선정

변수선택 | SHAP | 욕설포함확률

SHAP Dot Plot



'욕설포함확률'은 FI, SHAP에서 모두 높은 중요도를 보였기 때문에 **추가변수**로 선정

4. 변수 선택 및 모델 선정

변수선택 | SHAP | 욕설포함확률

개별 행의 Shapely Value에 대한 설명을 제공하는 SHAP Force Plot으로 확인한 결과

"욕설포함확률와 고소확률에 양의 관계가 있을 것이다"라는 가설과 상통하게 도출됨

Train data의 79번째 행

"여기저기 대주고 술얻어쳐먹고 밥얻어먹고 니가 몸파는년보다 더 더러운년이야 꺼져 미친년아"

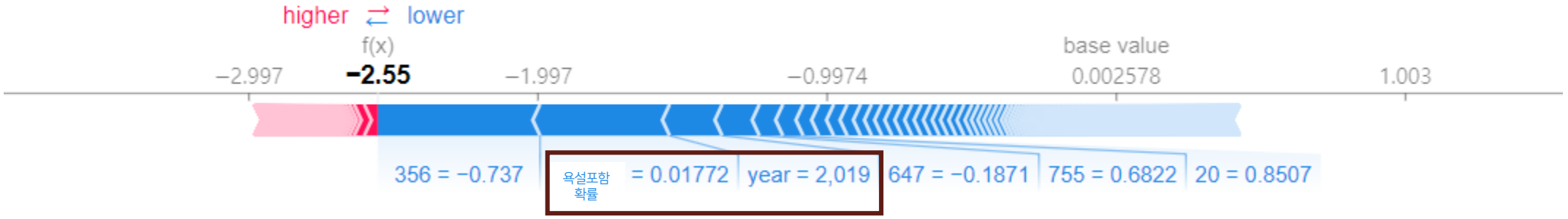
'욕설포함확률' 값이 매우 높아서 1로 예측되는 경우를 설명하는 force plot

4. 변수 선택 및 모델 선정

변수선택

Train data의 105번째 행

"겁나 웃기네 ㅋ"



'year'가 과거이고 '욕설포함확률' 값이 낮아서 고소여부가 0으로 예측되는 경우를 설명하는 force plot

4. 변수 선택 및 모델 선정

최종 데이터셋

변수선택까지 마친 진짜 찐 최종 데이터셋

	임베딩벡터		임베딩벡터	해당연도	댓글 길이	label
1	0.470965	…	-1.895960	2016	16	1
2	0.956919	…	-0.369123	2017	58	1
3	1.005964	…	-1.503245	2017	74	1
…	…	…	…	…	…	1
589	0.713427	…	-0.65980	2021	23	1
590	0.412480	…	-1.345210	2021	7	0
591	0.399288	…	-1.161543	2020	6	0
…	…	…	…	…	…	0
…	…	…	…	…	…	0
1178	0.985804	…	-1.441116	2021	15	0

1행~589행 : 판례 댓글
580행~1178행 : 일반 댓글

5. 최종 모델링

5. 최종 모델링

모델 성능 평가 지표

Custom Score

분류모델의 성능을 평가하는 다양한 metric 중

FP를 반영하는 정밀도와 **FN을 반영하는 재현율** 고려

댓글작성자에게 **경각심을 주기 위해 정밀도**를,

실제 클래스 비율을 반영하기 위해 재현율을 고려하기로 함

$$Custom\_score = (a \times FN) + (b \times FP)$$

FN과 FP의 반영비율에 대한 적절한 조정이 필요!

5. 최종 모델링

모델 성능 평가 지표

Custom Score

"댓글고소비율"에 관한 선행연구, 통계자료 전무

"댓글구속비율" 을 클래스 비율로 도입할 것을 고려해 보았으나

이는 클래스 비율 0.06%(60,000 : 1) 수준으로,

Custom Score의 가중치로서 부적합하다고 판단.

데이터셋의 **클래스 비율을 60,000 : 1로 재구성 하는 것 역시 학습률 저해가 우려됨**

본 연구의 목적은 **FN과 FP의 균형**을 유지하는 모델의 개발이므로

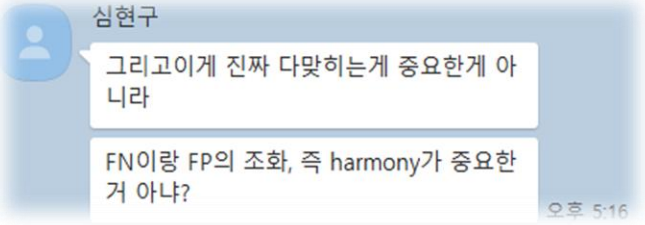
$$\text{Custom\_score} = (a \times FN) + (b \times FP)$$

FN과 FP의 조화평균을 이용한 **F1 Score**를 최종 지표로 활용하기로 결정함

FN과 FP의 **반영비율**에 대한 적절한 조정이 필요!

5. 최종 모델링

모델 성능 평가 지표



F1 Score 사용 이유

F1_score는 FP와 FN 모두를 고려하며,

특히 **실제 양성 및 음성 샘플이 중요한 경우**에 유용하므로

모델링 목적에 적합하다고 판단

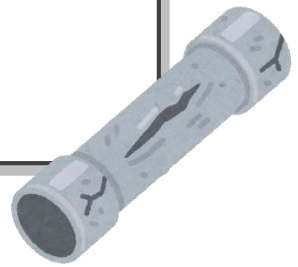
$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

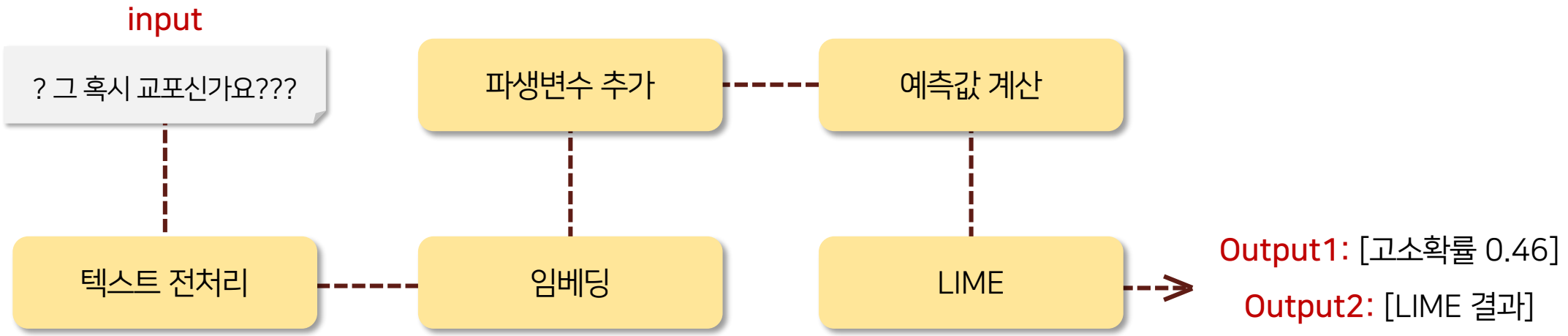
5. 최종 모델링

모델링 파이프라인

파이프라이닝 (Pipelining)

sklearn에서 제공하는 라이브러리로, 흩어져 있는 다중 프로세스를
한 번의 input과 한 번의 output으로 묶어서 처리하는 방법





5. 최종 모델링

모델링 파이프라인 | 텍스트 전처리

앞서 최종모델 선정 과정에서 진행한 전처리와 같은 과정을

댓글작성자로부터 입력받는 댓글에도 똑같이 적용

Input	Psat주1제분석 드디어 끝 ^^!!ㅎㅎㅎㅎㅎ
동일 음운 반복 축약	Psat주1제분석 드디어 끝 ^^!!ㅎㅎ
숫자, 영어, 특수문자 제거	주제분석 드디어 끝 ㅎㅎ
Preprocessed Output	주제분석 드디어 끝 ㅎㅎ

5. 최종 모델링

모델링 파이프라인 | 임베딩

댓글 내용
“이 시발새끼야 넌 내가 죽인다”
“네 여자친구랑 섹스해도 돼?”
“보물섬 형들 너무 잼씀 레알루다가”
…

Vector0	vector1	…	vector767	vector768
0.75687	0.00012	…	0.89820	1.77873
0.33324	0.020203	…	2.9901	1.40274
1.8569	3.14922	…	3.2207	1.26649
…	…	…	…	…

전처리가 완료된 상태에서 사전학습된 KLUE-BERT모델로 임베딩벡터 추출

5. 최종 모델링

모델링 파이프라인 | 파생변수 추가

Vector0	vector1	···	vector767	vector768
0.75687	0.00012	···	0.89820	1.77873
0.33324	0.020203	···	2.9901	1.40274
1.8569	3.14922	···	3.2207	1.26649
...	...	...	...	...

+

year	fuckyou
2017	0.9888
2021	0.3726
2016	0.2899
...	...

원래댓글에서 파생된 변수인 **작성연도**와 **욕설포함여부**를 추출 및 계산하여

임베딩벡터에 결합

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

Vector0	vector1	…	vector768	year	fuckyou
0.75687	0.00012	…	1.77873	2017	0.9888
0.33324	0.020203	…	1.40274	2021	0.3726
1.8569	3.14922	…	1.26649	2016	0.2899
…	…	…	…	…	…

댓글 내용	고소여부
“이 시발새끼야 넌 내가 죽인다”	1
“네 여자친구랑 섹스해도 돼?”	1
“보물섬 형들 너무 잼씀 레알루다가”	0
…	…

완성된 데이터프레임을 최종 LGBMClassifier 분류모델에 넣고 고소여부 계산

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

확률값 계산 방법

활성화함수로 압축된 single value를 확률값으로 활용 고려

은닉층을 거쳐 최종출력층에서 두 값으로 압축 → [a, b]

a가 크면 0, b가 크면 1로 분류하는 프로세스

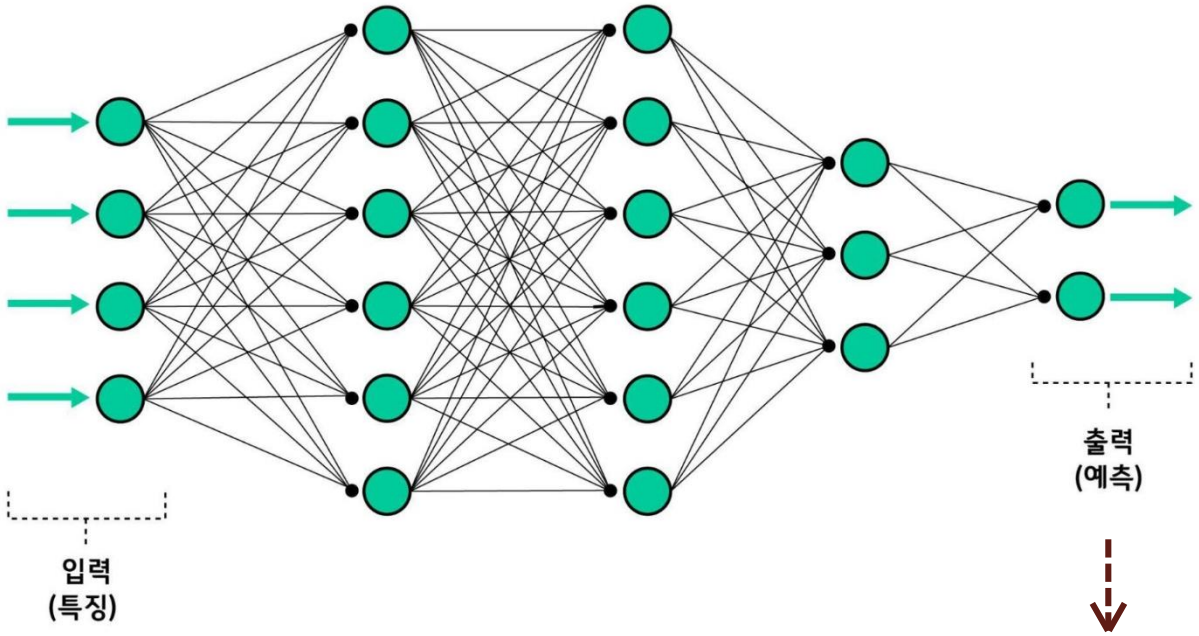
b값을 1이 될 확률, 즉 고소확률로 취급!!

우리는 고소여부에서 나아가 고소확률을 제시하기로 했으므로

최종적으로 **고소여부를 판별하기 전 상태**에서 **고소확률** 추출

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)



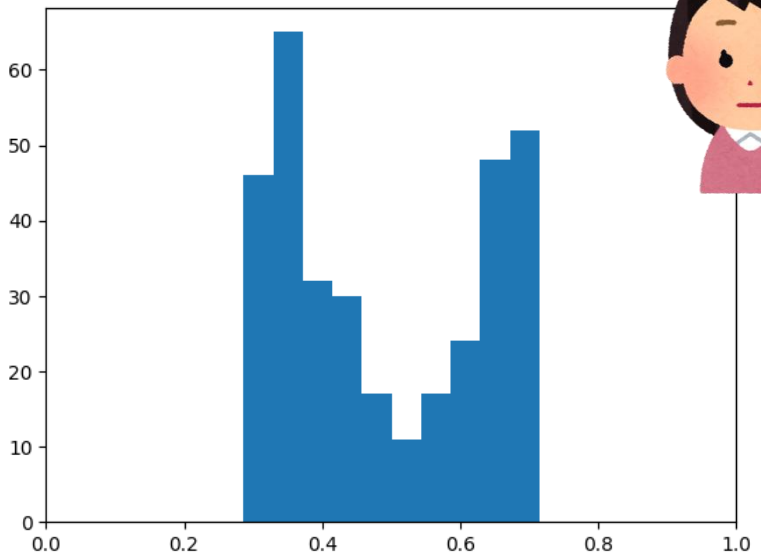
[a, b] = [0.34, 0.66]

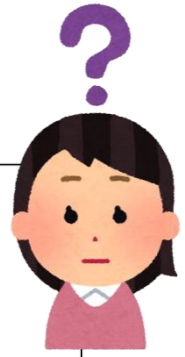
하지만 이 predict_proba() 결과를 고소확률로 바로 채택 시 문제가 발생할 수 있음!

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

Predict_proba() dist





확률값이 중앙으로 몰리는게 왜 문제인데?

댓글	고소확률
"여기저기 대주고 술얻어쳐먹고 밥얻어먹고 니가 몸파는년보다 더 더러운 년이야 꺼져 미친년아"	66.7147%
"아 짜증나네 죽을래?"	64.2339%

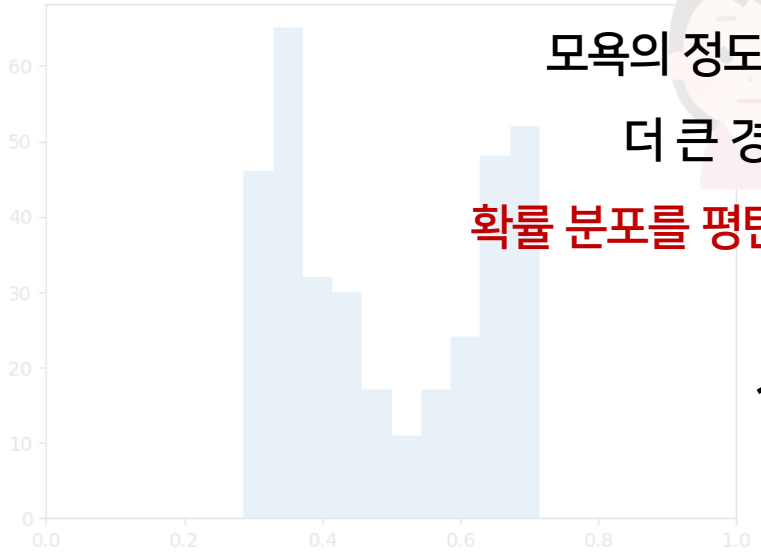
도출된 **확률 값이 중앙에 몰려 있는 문제** 발생

심한 모욕과 애매한 모욕의 고소확률이

매우 비슷하게 계산되는 결과

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)



도출된 확률 값이 중앙에 몰려 있는 문제 발생

모욕의 정도가 심할수록 댓글 작성자에게

더 큰 경각심을 심어줄 수 있도록

확률 분포를 평탄화 할 수 있는 방안을 논의 必 …

1. 백분율 재조정
2. Calibration

댓글	고소확률
“여기저기 대주고 술얻어쳐먹고 밥얻어먹고 니가 몸파는년보다 더 더러운 년이야 꺼져 미친년아”	66.7147%
“아 짜증나네 죽을래?”	64.2339%

심한 모욕과 애매한 모욕의 고소확률이

매우 비슷하게 계산되는 결과

5. 최종 모델링

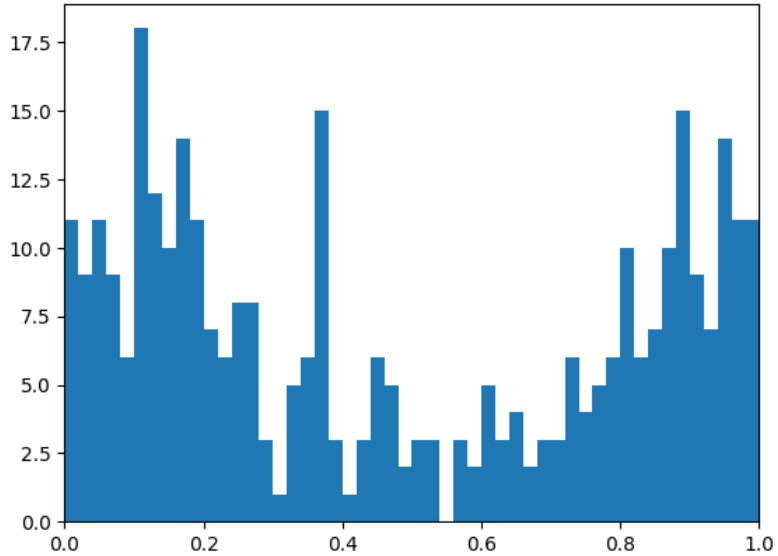
모델링 파이프라인 | 백분율 재조정

① 백분율 재조정

원래 분포의 최솟값을 0, 최댓값을 100으로 설정 후
그 사이의 값들을 percentile로 분산시킴

클린업 2주차 패키지과제에서 구간으로 설정된 범주형 변수를
백분율로 바꾸는 과정에서 아이디어를 얻음!
패키지 혐오를 멈춰주세요…

백분율 재조정 결과



5. 최종 모델링

모델링 파이프라인 | Calibration

② Calibration

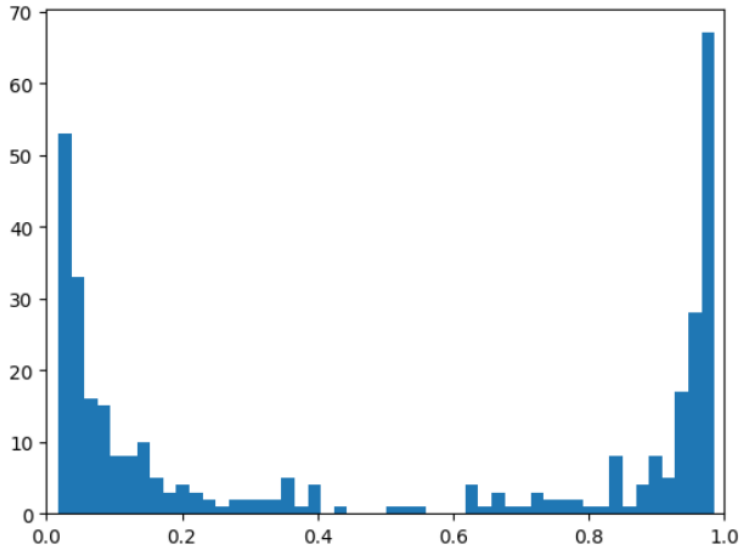
모형의 출력값이 **실제 확률을 반영하도록** 만드는 것

1) Histrogram Binning

2) Platt scaling

3) Temperature scaling

Platt Scaling 결과



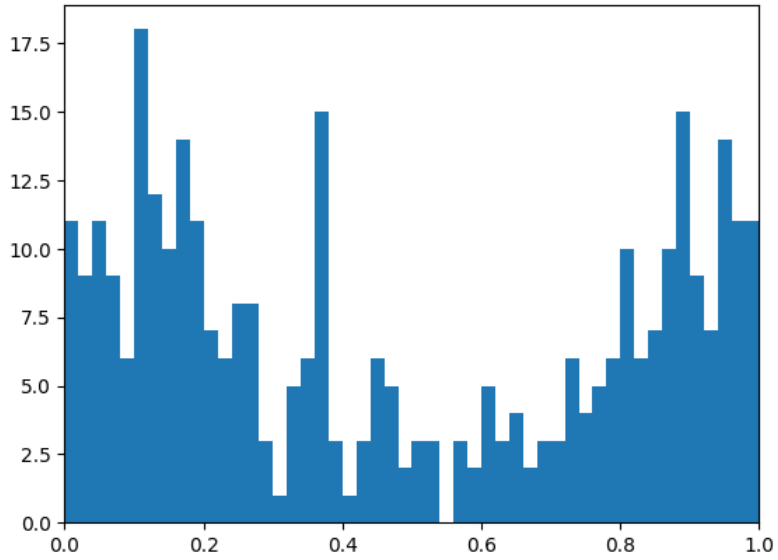
Calibration 적용 결과 고소 확률에 현실 확률 분포가 반영되어 0과 1에 집중됨을 확인.

그러나 다양한 확률값의 제공을 통해 **현실적인 수치로 경각심을 주려는 분석 목적과 어긋남**.

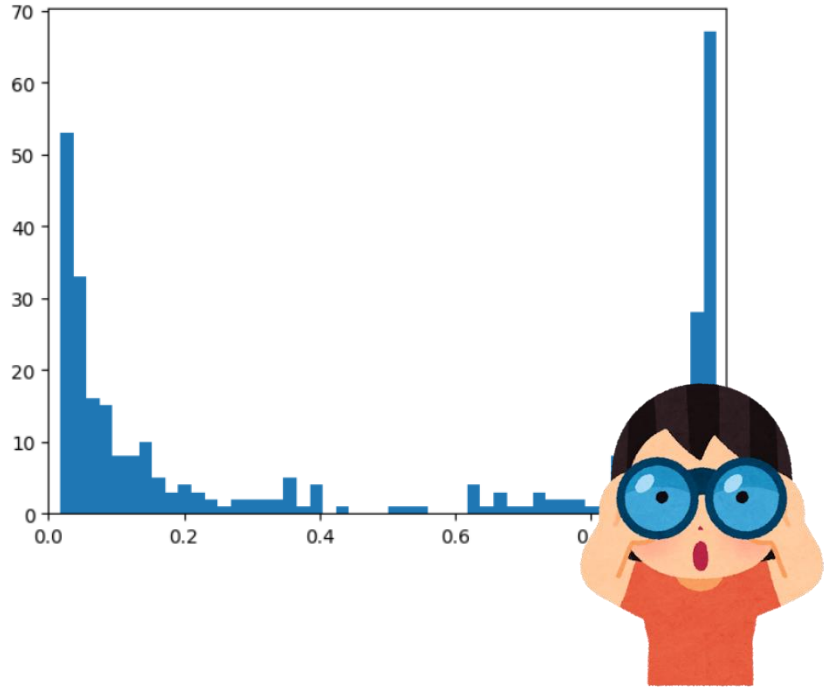
5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

백분율 재조정 결과



Platt Scaling 결과



더 smooth하게 scaling된 백분율 재조정 선택!

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

이렇게 변형을 가한 값이 0~1 사이의 값이라고 해서

'확률'이라고 표현하는 것이 타당한가를 놓고 논의하기도 함.

행동경제 학자

확률은 **합리적인 믿음**의 정도. 이것을 "**인지적 확률**" 이라고 한다!

황재홍, 2019. 불확실성에서의 의사결정과 확률: 케인즈와 행동경제학

이 개념을 차용해 '**확률**'이라는 표현을 유지하기로 결정.

더불어, 본 서비스는 **일반 대중을 이용 대상자**로 하기 때문에

'확률'이라는 일상적 표현을 넓은 의미에서 사용하여도 문제가 없음.

5. 최종 모델링

모델링 파이프라인 | 결과 예측 (분류)

행동경제학자 리처드 탈러의 실험

[상황 1]

걸리면 일주일 내에 무조건 죽게 되는 질병이 있습니다. 그런데 당신이 이 병에 걸릴 확률은 0.001%입니다. 이 병을 100%의 확률로 예방할 수 있는 백신이 있다면 얼마에 사시겠습니까?

[상황 2]

건강한 상태의 당신이 0.001%의 확률로 죽을 수 있는 의학 실험이 있습니다. 이 실험에 참여할 수 있다면 당신은 보상으로 얼마를 받아야 한다고 생각하십니까?

사람들은 [상황 1]에서는 200달러를,
[상황 2]에서는 10,000달러가 넘는 금액을 불렀음

손실 회피(Loss Aversion)

사람들이 실패와 손해를 본능적으로 회피하고
현재의 상태를 유지하기를 바라며 행동하는 경향

손실회피의 예시처럼
인지적 확률이 현실과 맞닿아 있음을
주제분석에 적용함으로써
"고소확률" 용어를 사용하기로 결론

5. 최종 모델링

모델링 파이프라인 | LIME

LIME(Locally Interpretable Model-agnostic Explanations)

특정 **예측 인스턴스 주변의 지역적(local) 설명을 생성**하여 모델의 동작을 해석하고 설명하는 기법

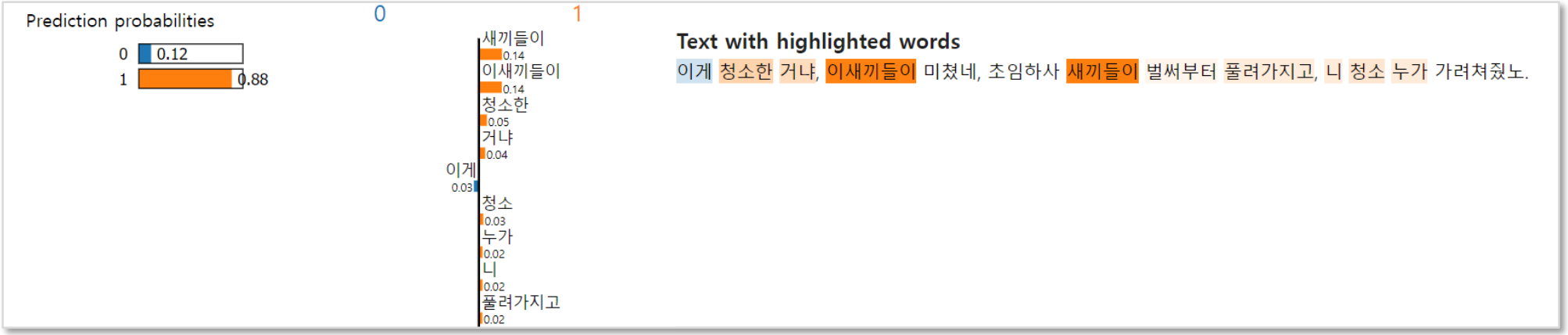
어떤 모델이든 사용할 수 있으며, 해당 모델의 내부 동작을 몰라도 모델의 예측을 설명하는 근사치를 만들 수 있음

LIME을 활용해 댓글을 구성하는 단어 중 **어떤 단어가 문제**가 되어

고소될 수 있다는 결론이 도출되었는지 **해석과 함께 시각화** 하고자 함

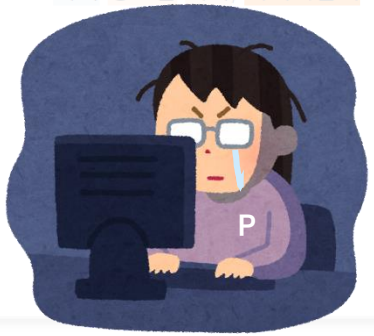
5. 최종 모델링

모델링 파이프라인 | LIME



5. 최종 모델링

모델링 파이프라인 | LIME



예원

이와 같은 결과창을 웹에 탑재해 뉴스 댓글 작성 후 등록 전에

고소확률과 함께 제시할 예정!

6. 결론

6. 결론

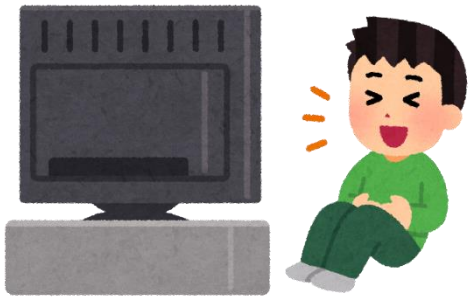
모델 예측 결과 | 실제로 고소당한 댓글

57.1 %	네 여자친구랑 섹스해도 되냐
70.1 %	이게 청소한 거냐 이새끼들이 미쳤네 초임하사 새끼들이 벌써부터 풀려가지고 니 청소 누가 가리처줬노
70.5%	이 미친놈아 너 나 지금 엿먹일려고 작정했냐 하라는 데만 하면 되는데 왜 이따위로 하냐 사람 말이 말 같지 않냐 하기 싫으면 때려치워 내가 할 테니 내가 괜히 너한테 시켰나 싶다 병들한테 부탁해서 걔들 보고 쓰라고 하는 게 더 빠르겠다

6. 결론

모델 예측 결과 | 일반 댓글

35.3%	형 나 면봉제발
28.4%	겁나 웃기넥
31.2%	동현이 형 연기하면 안 되겠다 개티남

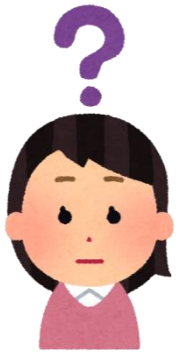


6. 결론

모델 예측 결과 | 그리고 …

고소확률이 낮게 나온 댓글에까지 경고를 제시할 경우
선플임에도 고소될 수도 있다는 점이 제시되어
모델 신뢰성이 하락하고 오히려 표현의 자유를 해칠 우려 발생

"와 정말 예쁜 멍멍이네요~ 한 번 보고 싶어요!"
[system] 당신의 고소확률은 16%입니다.



이게 고소당할 수도 있다고…?

악플러에게 경각심을 심어주려는 원래 목적에 맞게
고소확률이 0.5 이상으로 높게 예측되는 경우만 경고를 출력하기로 결정!

6. 결론

기대효과 ①

‘악플 노출량’이 아닌 ‘악플 생성량’ 자체를 감소

댓글 작성 과정에서 **고소 확률**을 제시해

댓글작성자가 **자발적으로 댓글 등록을 포기**하게 만듦으로써

악플 자체를 차단하는 기존의 유사한 서비스와 달리

표현의 자유를 보장하며 악플 생성을 줄일 수 있음

6. 결론

기대효과 ②

불필요한 사회적 비용을 절감

연세대학교 바른ICT연구소에 따르면 악플로 인한 사회경제적 비용은 연간 35조에 이름.

본 서비스를 도입할 경우 **악플 생성량이 감소**하며,

한정된 법자원의 효율적 분배로 사회 전체의 효용이 증가할 것으로 기대

6. 결론

기대효과 ③

건전한 인터넷 문화 형성에 기여

인터넷의 발달로 언제 어디서든 자신의 의견을 표현하는 것이 수월해진만큼
작성자는 자신의 댓글에 대한 책임을 간과하기 쉬움.
본 서비스를 통해 **모든 이용자**는 자신의 댓글에 대한 **책임감**을 가지게 되며,
건전한 인터넷 문화 형성에 기여할 수 있다.

6. 결론

기대효과 ④

뉴스 댓글뿐만 아니라 다른 분야에도 적용 가능

웹과의 연동을 통해 소셜 네트워크 서비스(SNS), **유튜브 댓글** 등에도 적용이 가능.

고소 원인을 제시하는 기능을 활용하면 댓글이 아닌 **게시글, 출판물**에도 적용이 가능.

6. 결론

기대효과 ⑤

시대의 흐름을 반영하여 사회문제 해결에 기여

웹에 탑재한 모델에 뉴스 데이터를 통해 사회문제와 현안을 반영할 수 있는 기능을 추가

→ **새롭게 발생하는 사회문제 역시 즉각적으로 반영**할 수 있을 것

피셋 활동 소감

다연

아직도 면접 보던 날이 생생한데 벌써 피셋 활동이 끝났네요! 아무것도 모르고 들어왔었는데 좋은 사람들을 많이 만나 1년 동안 정말 많은 것을 배우고 성장할 수 있었습니다~!!
시계열팀에서의 첫 학기는 정말 떼굴떼굴 구르면서 하나하나 배웠던 것 같아요..ㅎㅎ 지난 학기는 기존도 2명이라 힘들었을 텐데 신입들 배려해가며 잘 이끌어준 민이, 수린언니에게 항상 감사한 마음뿐입니다!!ㅜㅠ 그리구 기존 같은 신입이었던 동환오빠랑 유진이!! 같은 팀으로 활동할 수 있어서 정말정말 영광이었고 많이 배웠다는 말을 꼬옥~ 전하고 싶습니다 ㅎㅎ 1년동안 챙겨주고 함께해줘서 고마웠어용~~팀은 갈라졌어도 영원히 짱시야 짱시..ㅎㅎ
민, 수린언니, 동환오빠, 유진이의 그늘에서 벗어나 팀장으로 활동한 2학기 역시 팀원들에게 많은 것을 배운 시간이었습니다! 매일 "우리 가게 정상영업합니다"를 외치며 하와수일 뿐이라고 주장하지만, 한 학기 내내 정말 많은 것을 해준 예원이랑 현구오빠에게도 고생했고 고맙다는 말 전하고 싶습니다! 이렇게 능력있는 사람들과 함께했다는 사실만으로 든든했고, 주분도 잘 마무리할 수 있었습니다. 2학기 활동하느라 수고했고 이제 아프지 말자 이 올빼미들아!!! 다음으로 지난 학기 저와는 굉장히 비교될 만큼 멋진 신입이었던 동기오빠, 세인오빠도 정말 고생하셨습니다 ㅎㅎ 첫 학기부터 자연어 처리를 하게 되어서 정신 없었을 텐데 맡긴 일 척척 해내줘서 고맙습니당 다음 학기 걱정 없이 떠날 수 있을 것 같아요~!
마지막으로 바빴을 텐데 클린업, 주분 모두 도와주신 학회장팀에게도 무한 감사드립니다!! 31기, 32기 모두 수고 많으셨습니다. 항상 응원할게요~~

예원

피셋을 하기 전 피셋 하는 친구들이 항상 죽을 것 같지만 배우는 건 많아간다고 말했던 기억이 있는데 정말 맞는 말 같아요. 매일매일 관짝에 못을 박는 기분으로 명륜과 율전을 와리가리했지만 그만큼 배우는게 많았던 1년이라고 생각합니다. 부담스러울정도로 뛰어난 열정과 두뇌로 나를 공포에 떨게했던 회귀팀부터, 내가 묻어가려고 하면 티 안나게 묻어주던 6인분같은 5인분 시계열팀까지 모두 많은 도움을 주셔서 너무너무 감사하다는 말을 전하고 싶습니다!! 생각하지 못한 인연들을 학회에서 많이 챙겨가는 것 같아서 뿌듯하네요. 저는 이제 피셋을 떠나지만 ㅎㅎ 남아있는 신입 기수분들도 기존으로 활동하면서 의미있는 한 학기 보내시길 바랄게요 ~ !!

피셋 활동 소감

투머치토커 현구

분명 범주팀 합격했다고 전화받았었고, 분명 "내가… 기존…?" 하며 시계열팀에 (자발적으로) 던져졌는데 어느새 정신없이 두 학기가 끝나버렸네요… 다들 쓰는 멘트를 한번 써보도록 하겠습니다.

1. 스윗범주 지훈은선아 영원한 나의 멘토들. 너희 만날 때마다 선글라스 안챙겨서 시력을 잃게 생겼다 그저 빛… 많이 배웠다는 말은 너무 많이 해서 질렸지? 최고의 팀원이면서 인생친구들 얻은 것 같아서 항상 기쁘다ㅠ 희나야 이번학기 너무 바빠서 자주는 못봤지만(대용량 수업 맨날안가서 미안…^^) 넌 모르겠지만 개힘들때마다 네게 의지했어 시계열팀인데 시계열패키지도 못 도와주는 못난 기존프렌드였지만 이해해 줄거라 믿어^^ 4명팀에서 엄청난 주분 해내느라 고생 진짜 많았다 함께 성장한 사람으로서 너무 뿌듯하다 수고했어~~ 스윗범주 연말파티 합시다!

2. 짱시계열 다연예원아 그냥 내가 돈벌면 오쏘몰 한다발씩 갖다줄게 고생 너무 많았다 나만 바쁜 거 아닌데도 편의 봐주고 응원해줘서 너무 고마웠어! 아프다고 챙겨줄 때도 너무 감동이었다…bb 한 학기 같이 활동해보니까 너희가 어딜 가든 잘 할 이 시대의 **할머니**들이라는 것을 알게 되었다. 나는 이 학교를 떠나지만 배움에는 나이가 없으니… 할머니~ 이번이 마지막이예요~!

3. 우리 금쪽같은 신입 세인이랑 동기… 걱정이 별로 안 될만큼 많이 도와주고 내가 신입일 때보다 훨씬 더 많은 역량을 가지고 들어왔으니 너희의 폭풍성장은 안봐도 유튜브다~ 이말이야. 아니 패키지 뭐 별로 물어보지도 않고 맨날 척척내~~ 주분 내용 너무 복잡해서 지쳤을텐데 내색 안하고 해달라는거 성실하게 해줘서 정말 고마웠고 기존 활동 하면서 도움 필요한 거 있으면 언제든지 편하게 연락주세요 이 형님은 이제 "피셋 졸업생"이다ㅎㅎ 누가 2학기에 패키지함? 파이팅^^

4. 학회장팀 및 팀장님들… 총괄의 힘듦을 너무나 잘 알고 있어서 힘들어 보일 때는 안쓰러웠지만 항상 대단하다고 생각했습니다! 각 팀 잘 이끌어서 주분 결과도 한 팀도 빠짐 없이너무 흥미롭고 도움되는 내용이었고 클린업으로부터도 많이 배워갑니다 감사했어요~~~ 기존들도 다들 고생 많았습니다 다들 갓생 살다가 홈커밍때 봐요^__^

5. 용앤리치: 내가 취업하면 99% 너희 덕이다… 1%는 내가 지하철타고 이동한거?ㅎ 아기단계에서 시작했는데 잘한다고 해주고 건설적인 토론도 하며 많이 배웠습니다. 런던에서 보자구~~!!

피셋 활동 소감

세인

처음 공부해본 시계열 클린업부터 악명 높던 주제분석까지 신입 기수로 벌써 한학기를 마쳤습니다. 정말 겉핥기 수준으로 다뤘던 R과 파이썬을 통한 패키지 풀이, 클린업 세미나를 통한 통계적 학문 공부 그리고 이론을 실습으로 옮겨 본 주제분석까지 피셋에 들어오지 않았다면 경험해보지 못했을 정말 바쁜 한학기였습니다. 정말 힘들었지만 괴롭지 않게 활동을 마치게 도와준 팀장 및 팀원분들께 이 자리를 빌려 감사 인사를 전합니다.
들어본 과목이라고는 통원,통수,행대 밖에 없었지만 기존 기수들과 같은 신입이었던 동기 덕분에 한학기 무사히 활동했어. 팀장으로 갖은 일의 준비를 도맡고 매번 고생해준 다연이, 어떤 질문에도 항상 성심성의껏 대답해주고 친절하게 알려준 현구형, 패키지에 도움 많이 주고 항상 코드를 먼저 돌려보고 준비해주는 예원이, 같은 신입이라는 생각이 들지 않을 정도로 매사에 적극적이고 도움을 많이 받은 동기까지 정말 부족하기만 한 신입인 내가 한학기 활동을 무사히 마무리한 건 시계열 팀으로 활동했기 때문이라고 생각해, 정말 한학기 동안 고마웠어!!!

동기

1학기 동안의 피샛 활동이 끝났네요. 팀원들 덕분에 많이 배우고 성장 할 수 있었던 시간이었습니다. 어려운 내용을 질문하면 언제나 대답해준 훌륭한 기존 부원들, 특히 팀장님한테 질문을 많이 했는데 정말 잘 대답해주어서 감동이었습니다. 그리고 항상 먼저 팀에 필요한 부분을 알려주는 현구형, 코딩에 관한 부분은 예원이만 믿었습니다. 너무 든든했던 기존 부원들 너무 고생많았고, 많이 배웠습니다. 그리고 모르는 부분이 있으면 같이 문제를 해결하며 같이 성장한 세인이 형까지 한 학기 동안 고생하셨습니다. 두서없이 적었지만, 결론은, 다음 학기에는 이번 학기 동안의 활동을 되새기며 기존 부원으로서 더 좋은 모습 보여드리겠습니다. 파이팅!

감사합니다