판례데이터를 활용한
뉴스 댓글 고소 확률 예측

시계열자료분석팀

장다연 심현구 천예원 윤세인 이동기

목 차

1. 연구 소개
2. 데이터 수집 및 클렌징
3. 자연어 전처리
4. 변수 추가
5. 변수선택 및 모델선정
6. 최종 모델링
7. 결론

1. 연구 소개

선정 배경 | 사회적 배경

2014-2022년
사이버 명예훼손의 발생, 검거 수가
눈에 띄게 증가

[불법콘텐츠범죄 발생 및 검거 현황]

구 분		불법콘텐츠범죄			
		소계	사이버도박	사이버 명예훼손모욕	기타
2014	발생	18,299	4,271	8,880	794
	검거	14,643	4,047	6,241	616
2015	발생	23,163	3,352	15,043	524
	검거	17,388	3,365	10,202	346
…	…	…	…	…	…
2021	발생	39,278	5,505	28,988	436
	검거	26,284	5,216	17,243	321
2022	발생	35,903	2,997	29,258	447
	검거	23,683	2,838	18,242	268

출처 : 경찰청 사이버수사

1. 연구 소개

선정 배경 | 정치적 배경

모욕죄

1년 이하 징역이나 금고 또는 2백만원 이하의 벌금

정보통신망법상 명예훼손죄

3년 이하 징역 또는 3천만원 이하 벌금형,
허위사실 유포 시 7년 이하의 징역
또는 5천만원 이하의 벌금

동아일보, 2023.07.05

순식간에 퍼지는 '악성 댓글'
규제 있지만 실제 처벌은 미미

*… 반면 악성 댓글에 대한 규제와 처벌은 미미하다는 지적이다. 징역형까지 가능한 법 규정과 달리 대부분 **기소유예나 벌금형에 그치고 있기 때문**이다. …*

현행법과 달리 실제 처벌은 경미하게 이루어짐

1. 연구 소개

선정 배경 | 정치적 배경

악성 댓글로 인한 피해 사례가 증가함에 따라 **징벌적 손해배상** 도입에 대한 필요성이 계속해서 언급되고 있으나

법안이 통과되기가 쉽지 않고, 법리 해석과 적용까지 시간이 오래 걸림

징벌적 손해배상

민사재판에서 가해자의 행위가 악의적이고 반사회적일 경우

실제 손해액보다 훨씬 더 많은 손해배상을 부과하는 제도

사후적 처벌의 측면에서는 거쳐야 할 관문이 많음

이는 데이터를 통해 해결할 수 없는 부분이라고 판단

사전적 예방안 중 데이터 분석으로

문제를 완화할 수 있는 방안에 주목!

1. 연구 소개

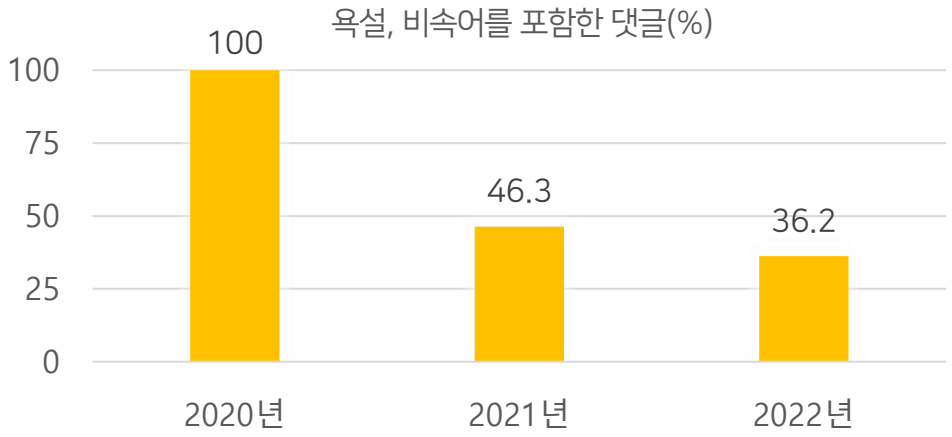
대응 현황

Kakao - 세이프봇

AI 기반 댓글 필터링 기능인 '세이프봇' 도입

욕설, 비속어를 음표로 치환

운영 정책을 위반해 불쾌감을 주는 댓글 삭제, 자동 신고

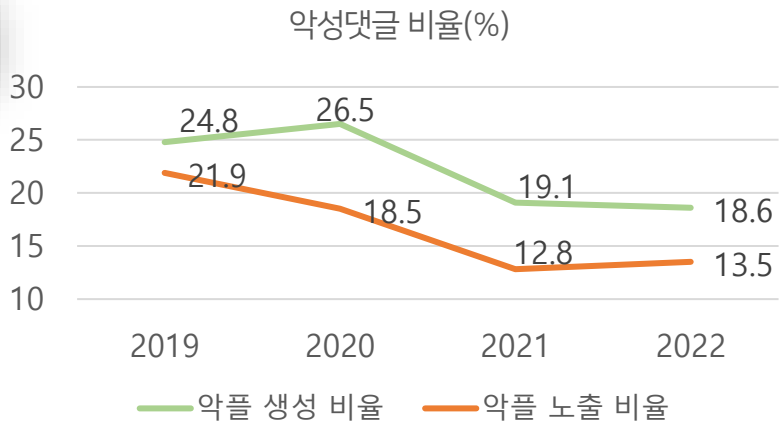


Naver - 클린봇

AI 기반 악성댓글 차단 프로그램인 'AI클린봇'이

욕설, 비하, 과도한 성적 표현을 포함하는 댓글 차단,

문장의 전체 맥락을 판단해 혐오표현 판단



세이프봇과 AI클린봇 도입 후 **악플 생성, 노출 비율 모두 감소하는 경향**

1. 연구 소개

대응 현황

Kakao - 세이프봇

AI 기반 댓글 필터링 기능인 '세이프봇' 도입

욕설,

운영 정책을 위반해

Naver – 클린봇

AI 기반 악성댓글

욕설, 비하, 과도한

문장의 전체 맥락을 판단해 혐오표현 판단

세이프봇과 AI클린봇 모두 악성 댓글을 삭제, 차단하는 방식

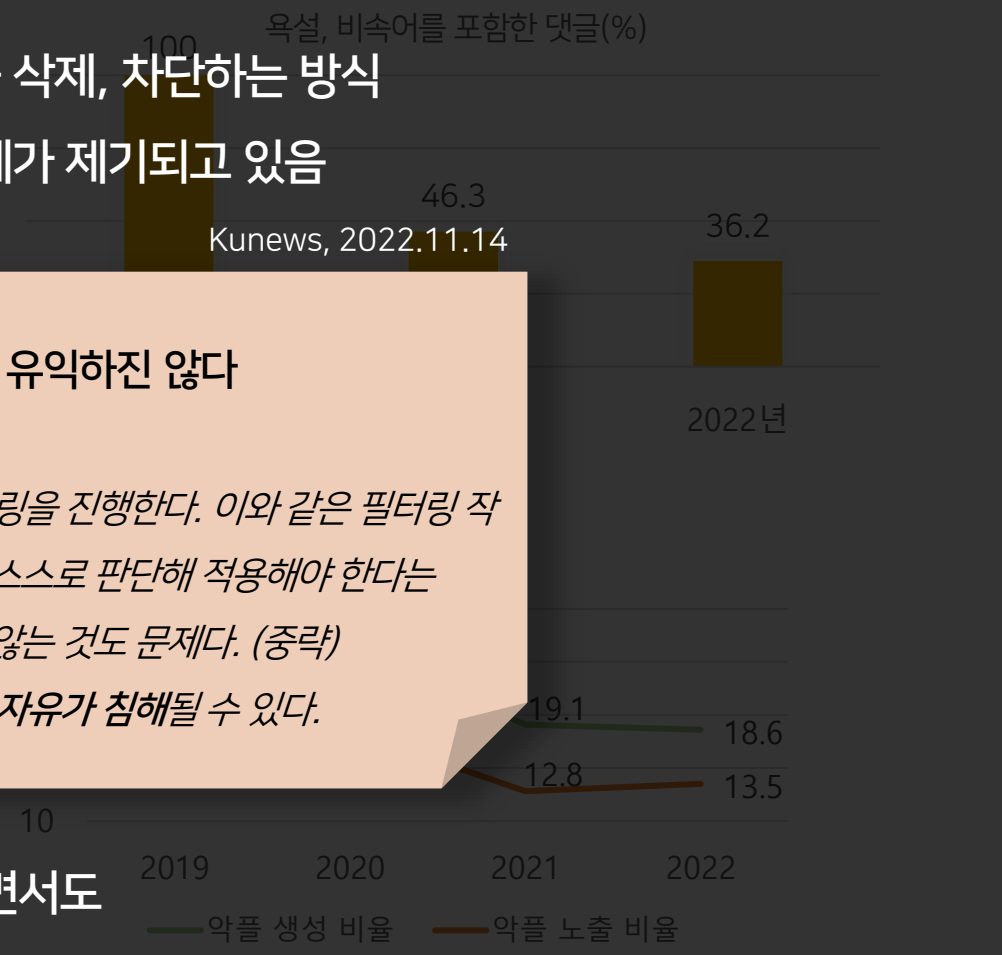
이는 **표현의 자유를 침해**한다는 문제가 제기되고 있음

Kunews, 2022.11.14

'클린봇'의 혐오 표현 필터링, 마냥 유익하지 않다

클린봇은 욕설 단어뿐 아니라 문장 맥락까지 고려한 필터링을 진행한다. 이와 같은 필터링 작업은 네이버와 같은 사기업이 표현의 자유의 한계를 스스로 판단해 적용해야 한다는 문제가 있다. 판단에 대한 공정성이 보장되지 않는 것도 문제다. (중략)

*이 과정에서 **소수의 견해가 필터링돼 표현의 자유가 침해**될 수 있다.*



세이프봇과 AI클린봇 도입 후 **악플 생성, 노출 비율 모두 감소하는 경향**

표현의 자유를 보장하면서도

작성자의 자발적 의지로 악플을 포기하게 만드는

악성 댓글 예방 서비스를 제공하는 것을 목표로 주제 선정

1. 연구 소개

연구 목적

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측

연구 목적 ① 표현의 자유를 보장하며 악플 문제 완화

악성 댓글 삭제가 아닌 고소 확률 제시를 통해
경각심을 주어 **자발적인** 악성 댓글 생성 방지

연구 목적 ② 건전한 인터넷 문화 형성

악플 작성자 뿐만 아니라 일반 대중에게도
댓글의 영향을 제시해 **건전한 인터넷 문화형성**에 기여

2. 데이터 수집 및 클렌징

판례 댓글데이터 수집 과정

판례 수집 과정

'로앤비' 사이트에서 키워드 검색을 통해

문제 댓글/발언이 명시된 판례전문 다운로드

"그 좆같은 새끼, 개같은 새끼, 쌍놈의 새끼라고 말하여 공연히 I를 모욕하였다"는 범죄사실에 대하여 **모욕죄**로 벌금 700,000원의 약식명령을 받았고, 위 약식명령은 2013 모욕죄로 벌금 700,000원의 약식명령을 받았으므로, 위 행위는 인사규정 제46호 제1항 제13호에서 정하는 징계사유에 해당한다(다만, 참가인이 이사장인 I에 대하여 위와 같이 모욕을 한 것만으로 이사장 선거에 개입하였다고 보기는 어렵고, 달리 이를 인정할 만한 증거가 없으므로, 이 부분에 관하여는 징계사유로 삼을 수 없다).

키워드별로 나누어 판례 다운로드

키워드	관련 판례 수
정보통신망이용촉진및정보보호등에 관한법률위반(명예훼손)	457
댓글	571
모욕죄	495
비방	3326
통신매체이용음란죄	37

4850개의 판례를 수집했으나

한 판례에 여러 키워드가 들어있는 경우 多

이를 중복데이터를 취급하여 제거 후

총 3,541개의 판례 확보

2. 데이터 수집 및 클렌징

일반 댓글데이터 수집 과정

일반 댓글데이터

AI HUB에서 **온라인 구어체 말뭉치** 다운로드

- 분야별로 구분된 json 파일
- 직접적인 반사회적용어, 비속어 등 masked

분야	내용	고소 label
유머	이때부터 살이조금씩 오르기시작하셨군요	0
유머	현석이형 턱살 레알 밥도둑	0
유머	(비속어)놈인가 죽여도 되는데 (반사회적용어) 안먹으면 안된다는건 뭔 (비속어)같은 (비속어)임	0
유머	참교육도 참교육인데 (반사회적용어)가 아깝게 느껴지네	0
방송	호동이 드릅게 답답하노	0
방송	진짜 저거 보면서 소스 왜이렇게 (이름)는(혐오표현) 생각함 보면서 불편	0
방송	여러분(혐오표현)망(이름)가(이름)피디들어오고나서망했습니다안그러면오래합니다	0
…	…	0

2. 데이터 수집 및 클렌징

데이터 클렌징

데이터 클렌징
댓글과 발언 위주로 데이터 수집 모델 학습에 사용 가능 여부가 불분명한 데이터에 대해 사용 가능한 기준 제시 후 선별

판례 댓글 클렌징
사실적시 및 허위사실 제거
심급 구별
…

일반 댓글 클렌징
다양한 분야의 댓글 중 방송, 유머 분야의 댓글 선택
판례 댓글과 다른 성격의 데이터 제거
…

3. 데이터 클렌징

불분명 판례 데이터 필터링

같은 댓글로 상소가 진행된 결과임에도 별개의 데이터로 취급한 이유

상급심 판례

상급심 판례의 경우 상소

여러 번 같은 판례 데이터가

중복처리해 제거하지 않

상소: 확정되지 않은 재판을 상급 법원에서

재판받을 수 있도록 하는 제도적 장치로,

항소, 상고, 항고 및 재항고를 포괄함

재판 경과

대법원 2020. 5. 28 선고 2019도12750 판결

23 선고 2019노721 판결

4 선고 2018고단452 판결

판결 [아동복지법위반 • 정보통신망이용 촉진

명예훼손)1 [공2020하, 1298]

있지 않다. '**학교폭력범**'이라는 단어는 '학교

접미사인 '범(3)'을 덧붙인 것으로서, '학교폭

고인은 '학교폭력범• 자체를 표현의 대상으로

상았을 뿐 특정인을 '학 하시키기에 충분한 구체적인 사실을 드러내 피해자의 명예를 훼손

하였다고 보아 이 사건 공소사실 중 정보통신망법 위반(명예훼손) 부분을 유죄로 판단하였

다. 원심판결 중 유죄 부분에는 정보통신…

상급심 판례

상급심 재판으로 갈 경우

재판 시간 경과, 추가 피고인 조사 등

피고인이 가질 부담이 증가

해당 사항을 반영하면 비슷한 텍스트에서의 고소확률이 증가해

악플 예방이라는 목적에 더 잘 맞게 학습이 이루어질 수 있을 것으로 판단

따라서, 심급에 따른 데이터를 별개로 취급

1심, 2심, 3심 선고에서 각각 추출한 데이터 : [학교폭력범]

2. 데이터 수집 및 클렌징

일반댓글 데이터

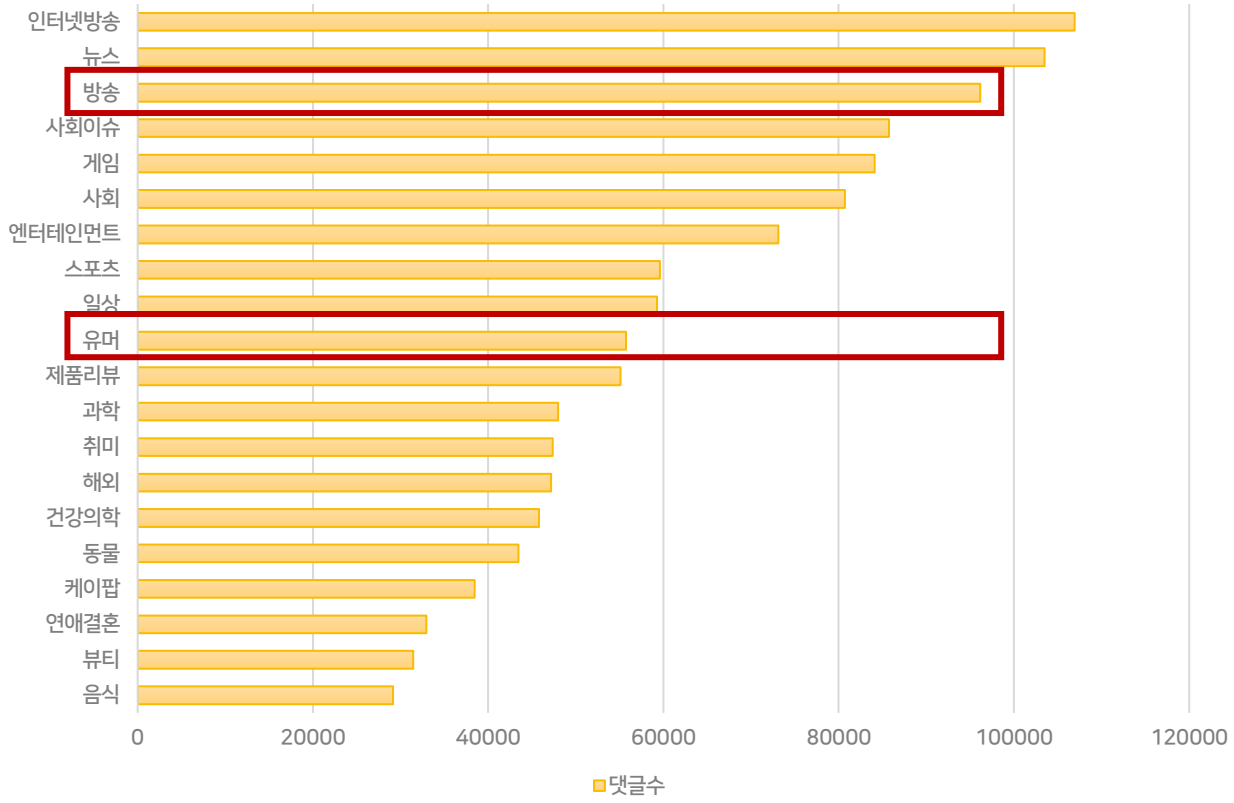
일반댓글 데이터

다양한 분야의 댓글 데이터 중
판례와 유사한 댓글을 가진 분야 선택

분야별 댓글 논조를 일일이 확인하며…

'취미'에는 진짜 취미에 진심인 댓글 뿐이네.

'과학'에는 감탄사만 가득해.



방송과 유머 댓글들이
학습용으로 적절하다고 판단!

2. 데이터 수집 및 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

2. 데이터 수집 및 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!

3. 자연어 전처리

토큰화(Tokenization)

토큰화 (Tokenization)

수집한 자연어 데이터를 '토큰(Token)'이라는 단위로 쪼개는 작업

토큰화로 나누어진 토큰은 자연어 처리 모델의 입력을 구성하는 기본 단위로서 작용

어떤 기준을 따라 토큰화를 하느냐에 따라서 모델의 성능이 달라질 수 있으므로

성능을 고도화 시킬 수 있는 적절한 토크나이저를 선택해야 함!

어절 단위 토큰화 | **형태소 단위 토큰화** | **Subword 기반 토큰화** | 음절 단위 토큰화

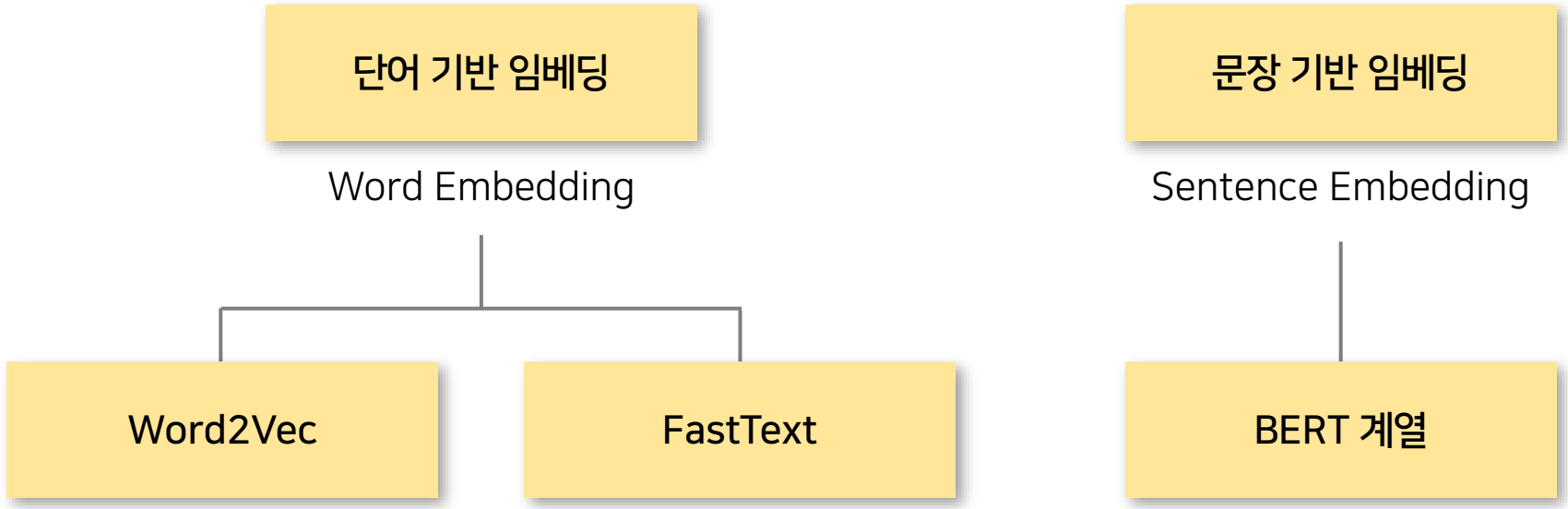
한국어의 특징을 살려 토큰화 하기 위한 토큰화 방법 선택

3. 자연어 전처리

임베딩 (Embedding)

임베딩 (Embedding)

자연어 문장 혹은 문서가 모델에 의해 처리될 수 있도록 벡터 형태로 변환하는 과정



3. 자연어 전처리

단어 기반 임베딩 | FastText

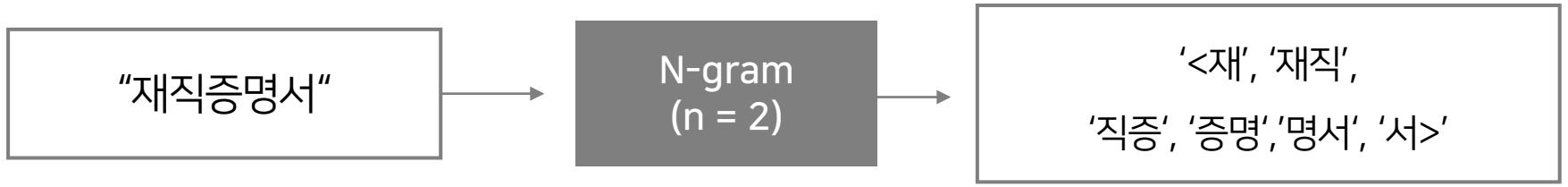
FastText

Word2Vec기법의 확장 매커니즘으로,

Word2Vec은 단어를 더 이상 쪼갤 수 없는 단어라고 간주하지만

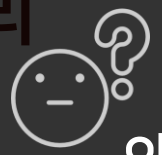
FastText는 하나의 단어 안에 여러 개의 단어가 결합되어 있는 것으로 간주한다.

이 때, subword의 추출을 위해 n-gram 기법을 사용한다.



이렇게 단어를 n개의 subword를 쪼개는 기법은 댓글 데이터에 효과적으로 작용할 수 있음

3. 자연어 전처리



왜 FastText가 댓글 데이터의 임베딩에 적합할까?

FastText

① 모르는 단어 (OOV)에 유연하게 대응 가능

모르는 단어 (OOV)가 등장하더라도 subword를 기준으로 다른 단어와의 유사도 계산 가능

➡ 신조어, 비속어가 많이 포함된 댓글 데이터 처리에 적합함!

② 희소한 단어 (Rare Word)에 유연하게 대응 가능

Word2Vec의 경우 등장 빈도수가 적은 단어에 대해 낮은 정확도를 보였음

그러나 FastText의 경우 해당 단어의 n-gram이 다른 단어의 n-gram과 겹친다면

Word2Vec과 비교하여 양호한 수준의 임베딩 벡터를 얻을 수 있음

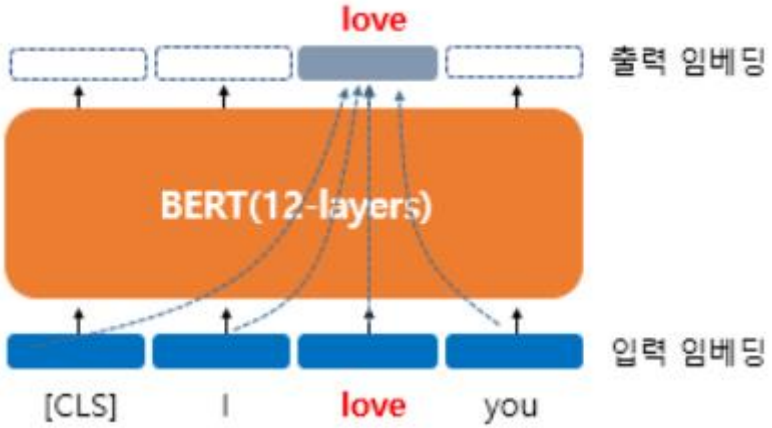
➡ 오타가 많은 댓글(희소 단어로 취급되는 단어)의 처리에 적합함!

3. 자연어 전처리

문장 기반 임베딩 | BERT

BERT 임베딩

대량의 단어 임베딩에 대해 사전 학습이 되어있는 임베딩 모델로,
기존 임베딩 모델이 맥락에 상관 없이 동일한 단어에 대해 동일한 임베딩을 반환했다면
BERT 임베딩은 **문맥을 반영**하여 문장을 임베딩함.



출력 임베딩의 'love'라는 단어를 임베딩 하기 위해
입력의 모든 단어 벡터들을 참고
⋮
주변 단어에 의해 동적으로 변화하는
중심 단어의 의미를 반영할 수 있음!

'모욕성 댓글'로 판단된 맥락을 학습시키기 위해 적합한 임베딩 방식!

4. 변수 추가

추가변수 생성

임베딩된 댓글 데이터 입력만으로

모델이 학습할 수 없다고 판단한 정보들을 변수로 추가

모델에 반영되지 못해 error가 된 정보량을 변수 추가를 통해

모델 변수로 편입시킴으로서 예측력 향상을 기대할 수 있음

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

텍스트에서 나온 결과들로 추가변수를 구성하는 경우,

임베딩 벡터에 **이미 이에 대한 정보가 반영되어 있을 가능성**

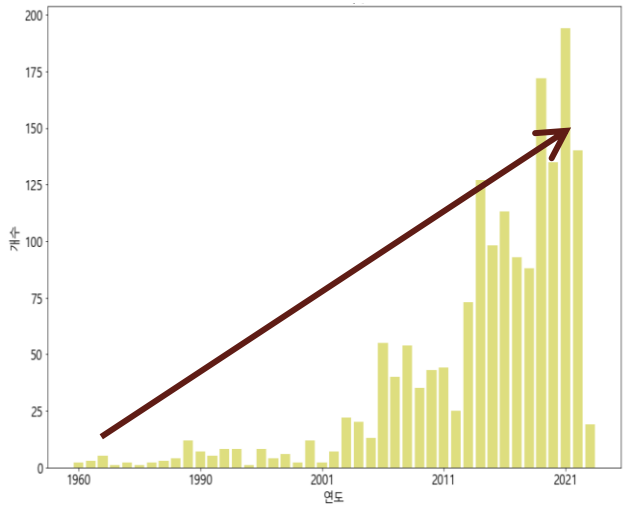
추가변수의 중요도가 임베딩 벡터의 중요도보다 낮을 경우,

해당 변수는 **제거**하기로 결정

4. 변수 추가

작성 년도 / 댓글 길이

연도별 판례 댓글 개수



최신판례는 댓글로 모욕한 경우의 범죄성립요건을
발언의 경우보다 완화하는 경향

["댓글에 `무뇌아` 단어 썼으면 모욕죄"]

김 씨는 지난 2013년 한 인터넷 카페에 게시된 글에 윤모씨를
무뇌아로 지칭하는 댓글을 달아 윤씨를 모욕한 혐의를 받고 있다.
(중략)
"같은 단어라도 댓글 같은 **짧은 글에서는 전체 맥락을 살피기 어려워**
상대방을 비난하는 단어를 쓰면 모욕죄가 될 개연성이 크다"

악플과 판례의 경향성을 반영하여
판결연도/작성연도를 변수로 추가

댓글 길이가 **매우 짧으면** 그 안에 **특정성**, **공연성**과 같은
범죄구성요소를 모두 포함하기 어려울 것이라고 판단

→

댓글 길이가 고소 확률에 유의미한 영향을 미칠 것이라는
가정 하에 **댓글의 길이**를 변수로 추가

4. 변수 추가

감성분석

AIhub에서 제공하는 한국어 감성 대화 말뭉치 학습데이터셋
(**기쁨, 슬픔, 분노, 불안, 당황, 상처**의 6가지 감정으로 라벨링이 완료된 데이터)을
KoBERT 모델에 학습시킨 후 다중분류

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

4. 변수 추가

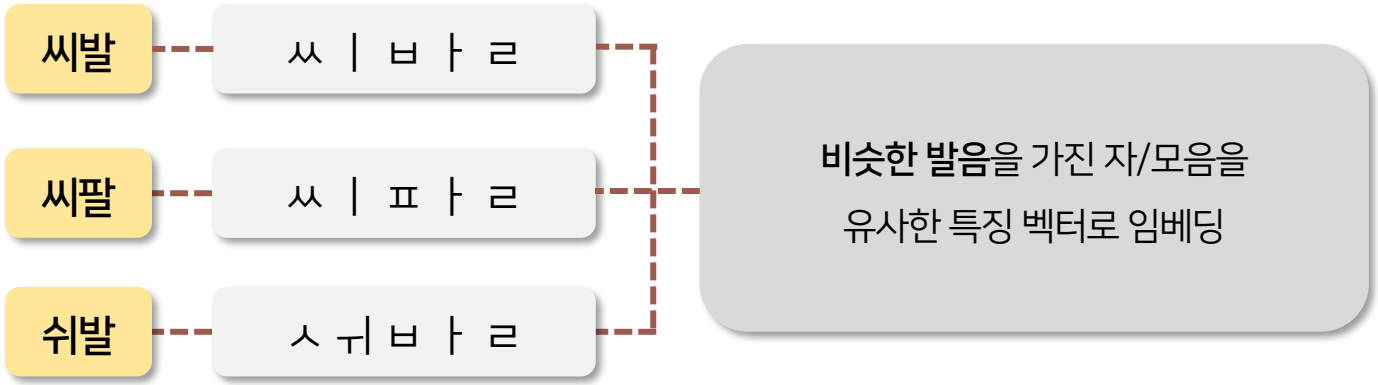
욕설 포함 확률

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

→ **MFCC 임베딩 기반** 모델로 댓글 내의 욕설 포함 확률을 탐지해 파생변수로 활용

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

5. 모델 선정 및 변수 선택

모델링 조합 선택

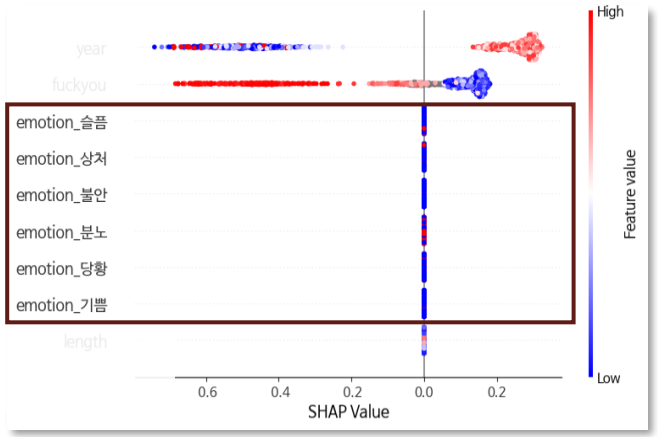
다양한 토큰화, 임베딩, 예측 모델 조합 중
최고성능을 낸 KLUE-BERT 임베딩 +
LightGBMClassifier 선택!

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

5. 모델 선정 및 변수 선택

변수선택 | SHAP

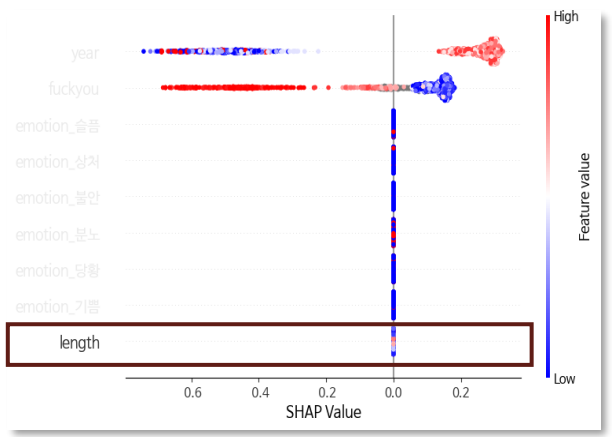
감정벡터 Dot Plot



감정벡터는 전체적으로 **낮은 변수중요도**

→ 임베딩벡터에 포함된 정보로 해석, 삭제

댓글길이 Dot Plot



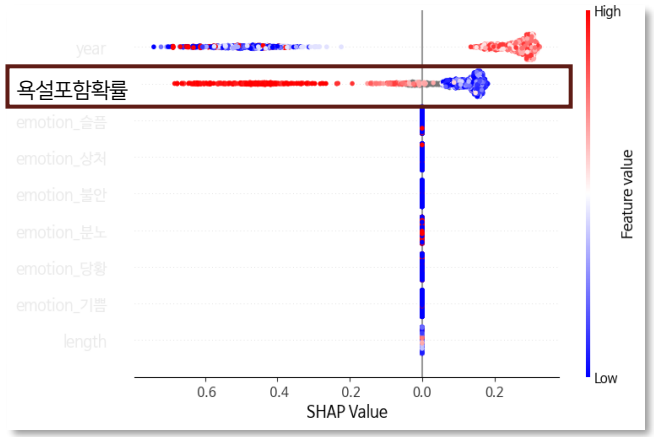
'댓글길이' 변수는 **중요도가 낮고** 추가 시 **성능이 하락**

판례댓글 중 "국민호텔녀" 등 짧은 댓글도

꽤 관측되기 때문이라고 해석,

학습 방해 변수로 판단하여 삭제

욕설포함확률 Dot Plot



'욕설포함확률'은 높은 중요도

영향을 미치는 방향도 기대와 상통하므로 **추가변수로 선정**

5. 모델 선정 및 변수 선택

변수선택 | 1. 모형 적합성 검정

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					

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

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

→ 0.1 유의수준 하에서 귀무가설을 기각해 full model 채택

추가로 year 변수에 대한 개별 유의성 검정을 진행함

5. 모델 선정 및 변수 선택

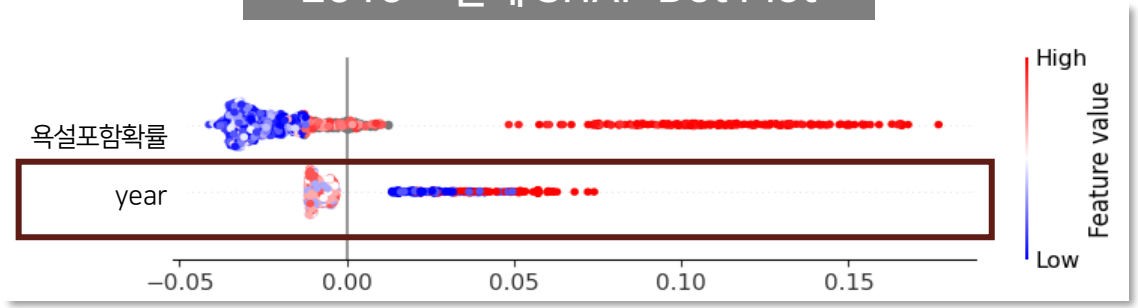
변수선택 | 2. 변수 유의성 검정

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 = 0.1316으로, 유의수준으로 설정한 0.1 경계에 존재함을 확인

섣불리 귀무가설을 채택/기각할 수 없다고 판단 → SHAP 결과 확인

2016 ~ 판례 SHAP Dot Plot



작성 연도와 고소여부가 양의 상관관계로 가설과 상통하는 결과 도출

'year'를 추가변수로 선정

5. 모델 선정 및 변수 선택

최종 데이터셋

변수선택까지 마친 최종 학습데이터셋

	임베딩벡터		임베딩벡터	해당연도	댓글길이	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

6. 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

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

FN과 FP의 반영비율에 대한 적절한 조정이 필요!

F1 Score

F1_score는 FP와 FN 모두를 고려하며,
특히 **실제 양성 및 음성 샘플이 중요한 경우**에 유용하므로
모델링 목적에 적합하다고 판단

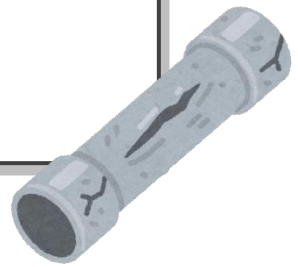
$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

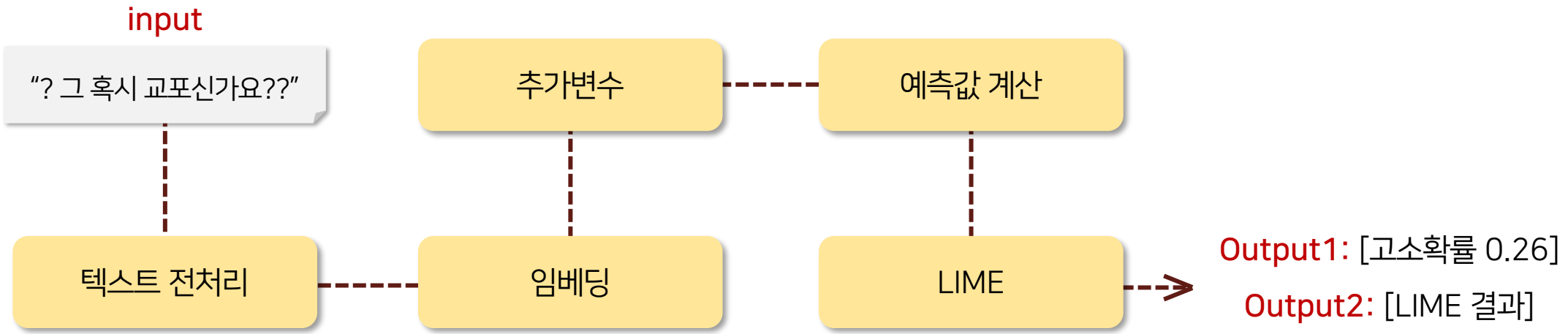
6. 최종 모델링

모델링 파이프라인

파이프라이닝 (Pipelining)

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





6. 최종 모델링

모델링 파이프라인 | 텍스트 전처리

앞서 데이터 클렌징 과정에서 진행한 전처리와 같은 과정을

댓글작성자가 입력하는 댓글에도 똑같이 적용

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

6. 최종 모델링

모델링 파이프라인 | 임베딩 + 추가변수

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
…	…	…	…	…

+

year	fuckyou
2017	0.9888
2021	0.3726
2016	0.2899
…	…

전처리가 완료된 상태에서 사전학습된 KLUE-BERT모델로 임베딩벡터 추출 후
원래댓글에서 파생된 변수인 **작성연도**와 **욕설포함여부**를 추출 및 계산하여 임베딩벡터에 결합

6. 최종 모델링

모델링 파이프라인 | 예측값 계산

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
…	…	…	…	…	…

LGBMClassifier →

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

본 연구는 고소여부에서 나아가 고소확률을 제시하기로 했으므로
최종적으로 **고소여부를 판별하기 전 상태**에서 **고소확률** 추출

확률값 계산 방법

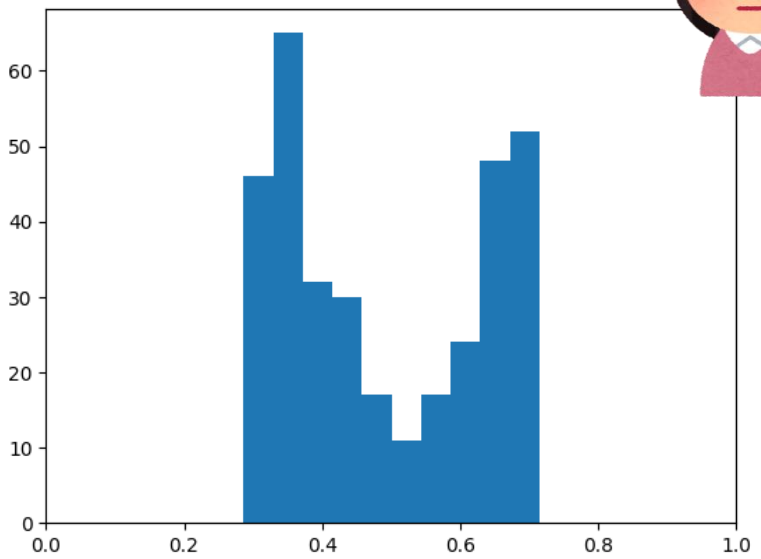
활성화함수로 압축된 single value(0~1)를 확률값으로 활용 고려
은닉층을 거쳐 최종출력층에서 두 값으로 압축 → [a, b]

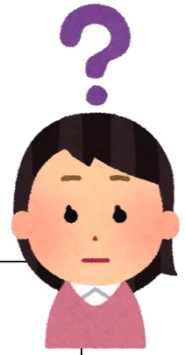
b값을 1이 될 확률, 즉 고소확률로 취급!!

6. 최종 모델링

모델링 파이프라인 | 예측값 계산

Predict_proba() dist





확률값이 중앙으로 몰리는게 왜 문제인데?

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

심한 모욕과 애매한 모욕의 고소확률이
매우 비슷하게 계산되는 결과

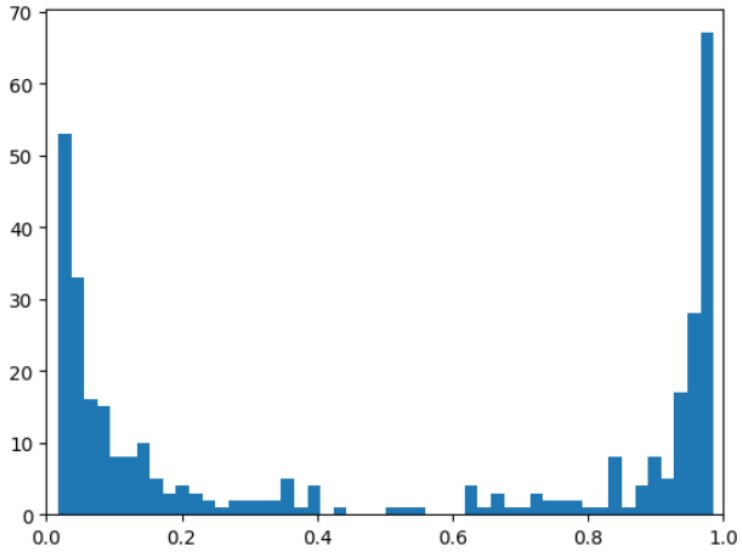
도출된 **확률 값이 중앙에 몰려 있는 문제** 발생

모욕의 정도가 심할수록 댓글 작성자에게
더 큰 경각심을 심어줄 수 있도록 분포 변형

6. 최종 모델링

모델링 파이프라인 | 예측값 계산

Platt Scaling 결과

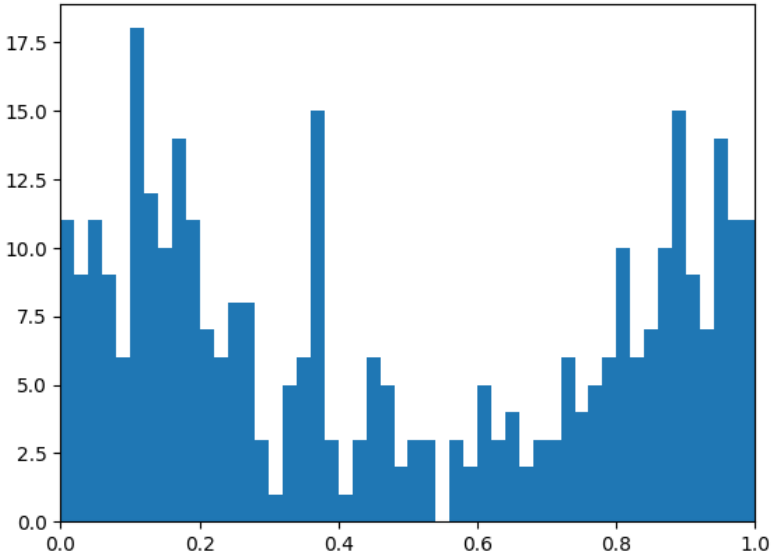


Calibration 적용 결과 고소 확률에 현실 확률 분포가 반영되어

0과 1에 집중됨을 확인.

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

백분율 재조정 결과



더 smooth하게 scaling된 **백분율 재조정 선택**!

6. 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**

7. 결론

고소 확률 예측 결과

고소 당한 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

일반댓글

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

7. 결론

고소 확률 예측 결과

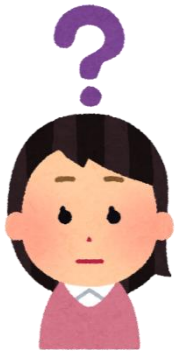
고소확률이 낮게 나온 댓글에까지 경고를 제시할 경우

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 목적에 맞게

고소확률이 0.5 이상으로 높게 예측되는 경우만 경고를 출력하기로 결정!

감사합니다