판례데이터를 활용한
뉴스 댓글 고소 확률 예측

시계열자료분석팀

장다연 심현구 천예원 윤세인 이동기

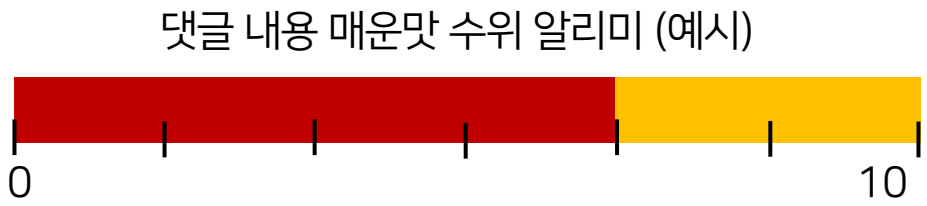
목 차

[주의]

개조심

실제 판례를 다루는 내용의 특성 상 매우 불쾌하고 선정적인 댓글 / 발언이 포함되어 있습니다.

각 챕터의 댓글 내용 매운맛 수위 알리미를 통해 마음의 준비를 한 뒤 읽어주세요.



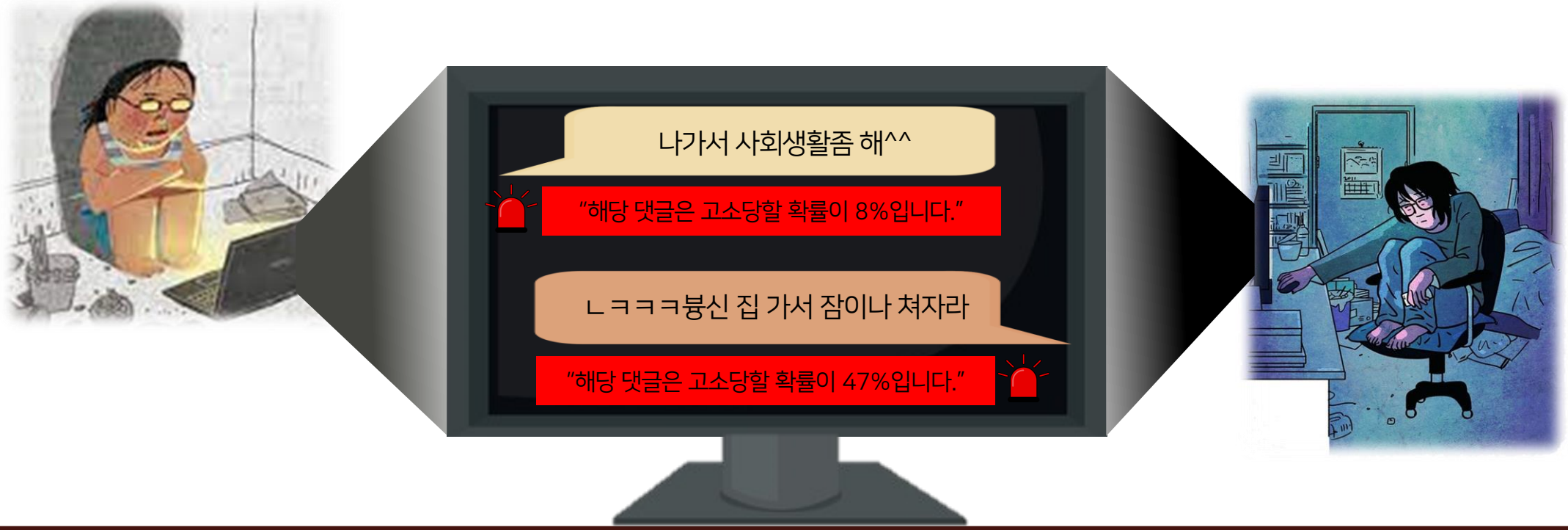
1. 연구 소개

댓글 내용 매운맛 수위 알리미

1. 연구 소개

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측



1. 연구 소개

선정 배경 | 사회적 배경

['모욕 댓글' 관련어 워드 클라우드]

From BIGKINDS, edited by 1000daughter

'모욕 댓글' 관련어로 만든 워드 클라우드를 통해

명예훼손, 온라인 커뮤니티 등이 가장 큰 문제임을 파악할 수 있음

1. 연구 소개

선정 배경 | 사회적 배경

2014-2022년 사이에
사이버 명예훼손의 발생, 검거 수가
눈에 띄게 증가했음을 알 수 있음

[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 | 사회적 배경

서울경제TV, 2023.08.23

악플 피해 법적 대응 증가세…
"무심코 작성한 악성 댓글로 전과자 낙인 "

23일 경찰청에 따르면 지난 2022년 사이버 명예훼손 및 모욕범죄 신고건수는 2만9,258건으로 역대 최대치를 기록했다. 2017년(1만3,348건)과 비교하면 5년 새 2배 이상 증가했다.

뉴시스, 2023.10.17

"악플 때문에…" 대중문화예술인 심리상담
전년比 4배 급증

지난해 한국콘텐츠진흥원이 운영하는 대중문화예술지원센터를 통해 심리상담을 지원 받은 대중문화예술인들이 전년 대비 4배 증가한 것으로 나타났다.

악성 댓글로 인해 사회 각 분야에서 댓글 작성자, 피해자 모두에게 악영향

1. 연구 소개

선정 배경 | 사회적 배경

사회적 비용

생산 주체가 부담하는 사적 비용과 재화 외에도
외부성으로 인한 사회적 부담 비용을 포함한 비용

세계일보, 2023.07.13.

'악플'에 매년 35조원 증발 …
"징벌적 배상제 도입을" 목소리도

연세대 바른ICT연구소에 따르면 악성 댓글로 인한 사회·경제적 비용은 연 35조3480억원에 이른다. (중략) 지난해 ***국내 GDP의 약 1.6%****에 달하는 막대한 사회적 비용이 발생한 셈이다.*

사회적 비용 측면에서도 악플은 해결 되어야 할 문제임

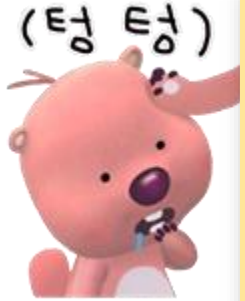
1. 연구 소개

선정 배경 | 사회적 배경

악성 댓글은 피해자를 극단적인 선택에까지 이르게 하고,

무심코 쓴 댓글이 고발당해 비용이 발생하는 등 심각한 사회문제로 근본적인 해결책이 필요함

문제 해결을 위해 시계열팀이 할 수 있는 일은?!



- 선플달기운동
- 강도 높은 처벌
- 삼청교육대

...

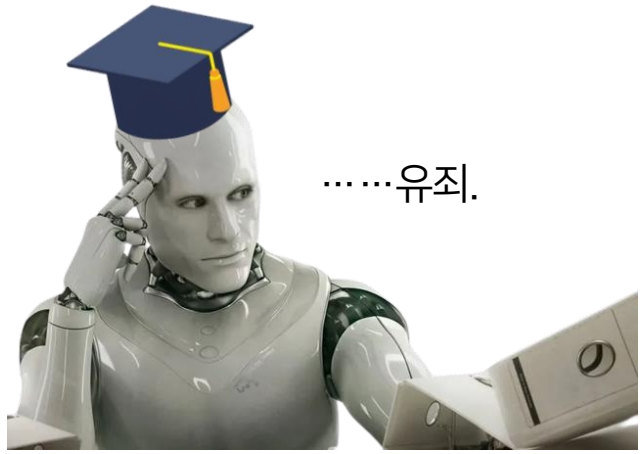
탁상공론 그만 !

피셋이라면

데이터를 통해 세상을 바라보자 !

1. 연구 소개

선정 배경 | 사회적 배경



중앙일보, 2023.06.26

"인간 판사와 79% 답 같았다"… 시간 아끼는 'AI 판사' 나올까

경기 부천시에서 서울 양천구까지 약 10㎞를 혈중알코올농도 0.182%(면허취소 이상) 상태로 운전한 B씨는 (중략) 이런 조건들을 AI에 입력하자 징역 10개월이 나왔다. 실제로 B씨는 징역 10개월을 선고받고 복역중이다. (중략)

*오세용 인천지법 부장판사는 "유사 사건을 검색해 사건별 양형 분포를 파악하는데 **시간·노력 절감** 효과가 있고, 신속하게 형량 범위를 판단할 수 있어 **복잡한 다른 쟁점에 집중할 수 있다**"는 점을 강조했다.*

법률 문제에 법관의 양형 보조자로서 AI의 도입 가능성이 제기됨

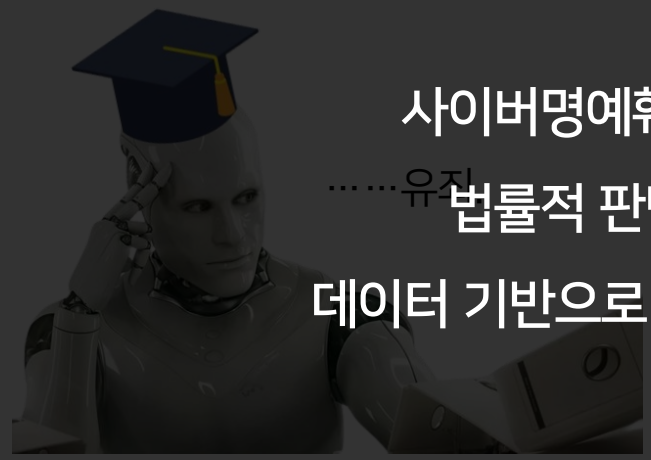
데이터를 학습한 모델을 활용해 법 문제를 해결할 수 있음

1. 연구 소개

선정 배경 | 사회적 배경

중앙일보, 2023.06.26

"인간 판사와 79% 답 같았다"… 시간 아끼는 'AI 판사' 나올까



AI 판사의 사례처럼

사이버명예훼손 발생 건수와 악플 피해 증가에 대한

……유죄

법률적 판단에 많은 사회적 비용이 드는 문제를

데이터 기반으로 완화할 수 있을 것이라는 아이디어를 얻음 !

(전략) 경기 구리시에서 서울 양천구까지 약 10㎞를 혈중알코올농도 0.182%(면허취소 0.1%) 상태로 운전한 B씨는 (중략) 이런 조건들을 AI에 입력하자 징역 10개월이 나왔다. 실제로 B씨는 징역 10개월을 선고받고

오세용 인천지법 부장판사는 "유사 사건을 검색해 사건별 양형 분포를 파악하는 [illegible]를 판단할 수 있어 복잡한 다른 쟁점에 집중할 수 있다"는 점을 강조했다. (후략)

법률 문제에 법관의 양형 보조자로서 AI의 도입 가능성이 제기됨

데이터를 학습한 모델을 활용해 법 문제를 해결할 수 있음

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

2023.07.05 동아일보

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 때문이다. …

현행법과 달리 실제 처벌이 강하게 이루어지지 않고 있음

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 처벌이 강하게 이루어지지 않고 있음

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

솜방망이 처벌 지긋지긋해!!!
징벌적 손해배상을 도입해야 한다!

VS

응 안돼 돌아가 ㅋㅋ

1. 연구 소개

선정 배경 | 정치적 배경

징벌적 손해배상

민사재판에서 가해자의 행위가 악의적이고 반사회적일 경우

실제 손해액보다 훨씬 더 많은 손해배상을 부과하는 제도

이렇듯 **사후적** 처벌 가중의 측면에서는 거쳐야 할 관문이 많음

이는 데이터를 통해 해결할 수 없는 부분이라고 판단

사전적 예방안 중 데이터 분석으로

문제를 완화할 수 있는 방안에 주목!

현황은 어떨까?

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

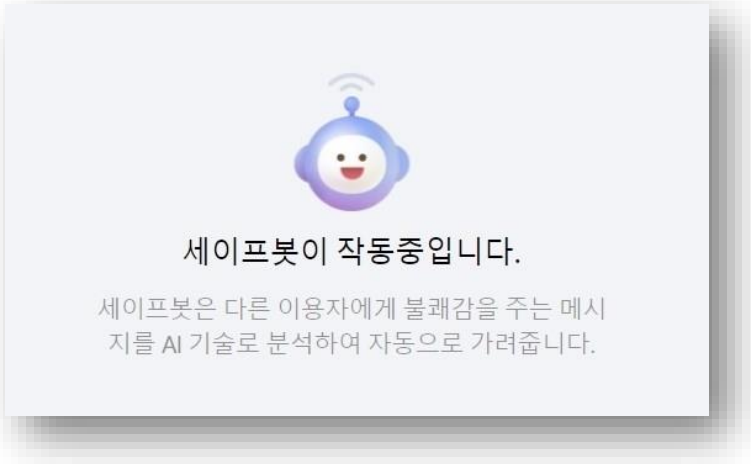
Naver – 클린봇

AI 기반 악성댓글 차단 프로그램인 'AI클린봇' 도입

지속적인 업데이트를 통해

욕설, 비하, 과도한 성적 표현을 포함하는 댓글 차단,

문장의 전체 맥락을 판단해 혐오표현 판단



1. 연구 소개

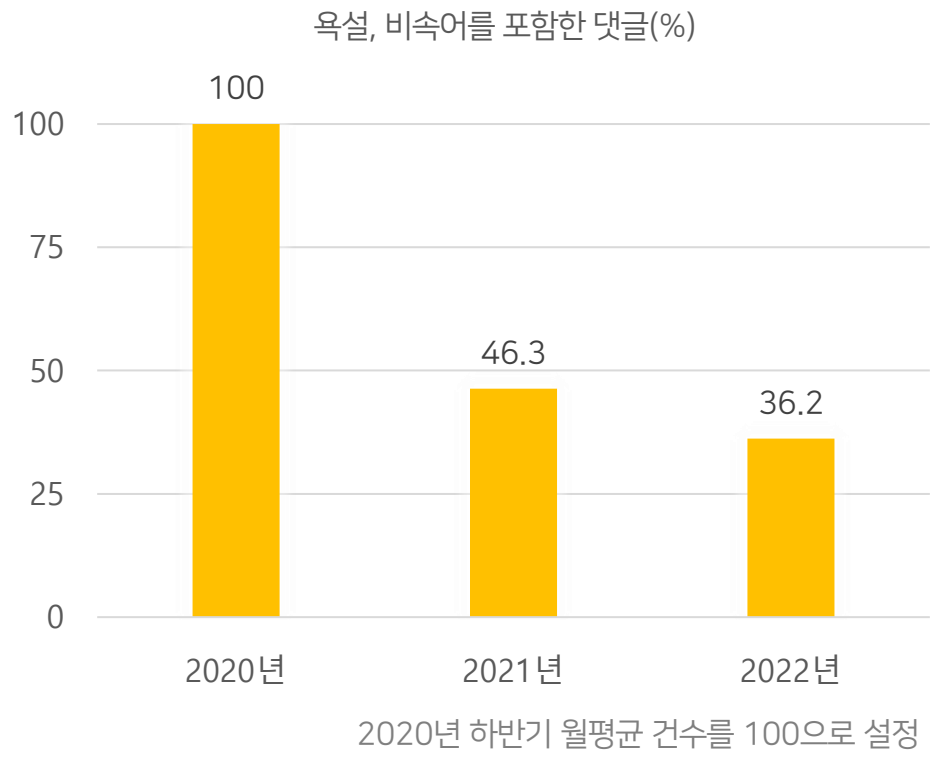
대응 현황

Kakao - 세이프봇

AI 기반 댓글 필터링 기능인 '세이프봇' 도입

욕설, 비속어를 음표로 치환

운영 정책을 위반해 불쾌감을 주는 댓글 삭제, 자동 신고



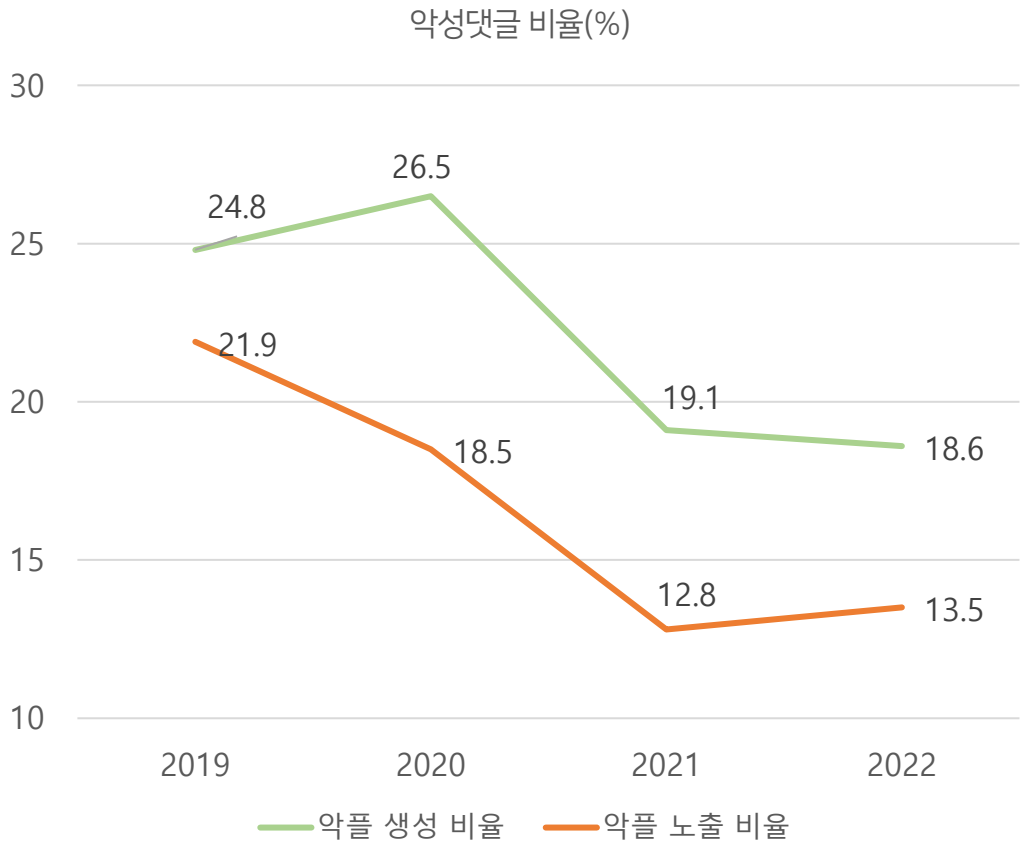
세이프봇 도입을 통해 **악성 댓글이 줄어들었음**

1. 연구 소개

대응 현황

Naver – 클린봇

AI 기반 악성댓글 차단 프로그램인 'AI클린봇'의
지속적인 업데이트를 통해
욕설, 비하, 과도한 성적 표현을 포함하는 댓글 차단,
문장의 전체 맥락을 판단해 혐오표현 판단



AI클린봇 도입을 통해 **악플 생성, 노출 비율 모두 낮아짐**

1. 연구 소개

대응 현황

세이프봇과 AI클린봇 모두 악성 댓글을 삭제, 차단하는 방식

이는 **표현의 자유를 침해**한다는 문제가 제기되고 있음

Kunews, 2022.11.14

Naver – 클린봇

AI 기반 악성댓글 차

지속적

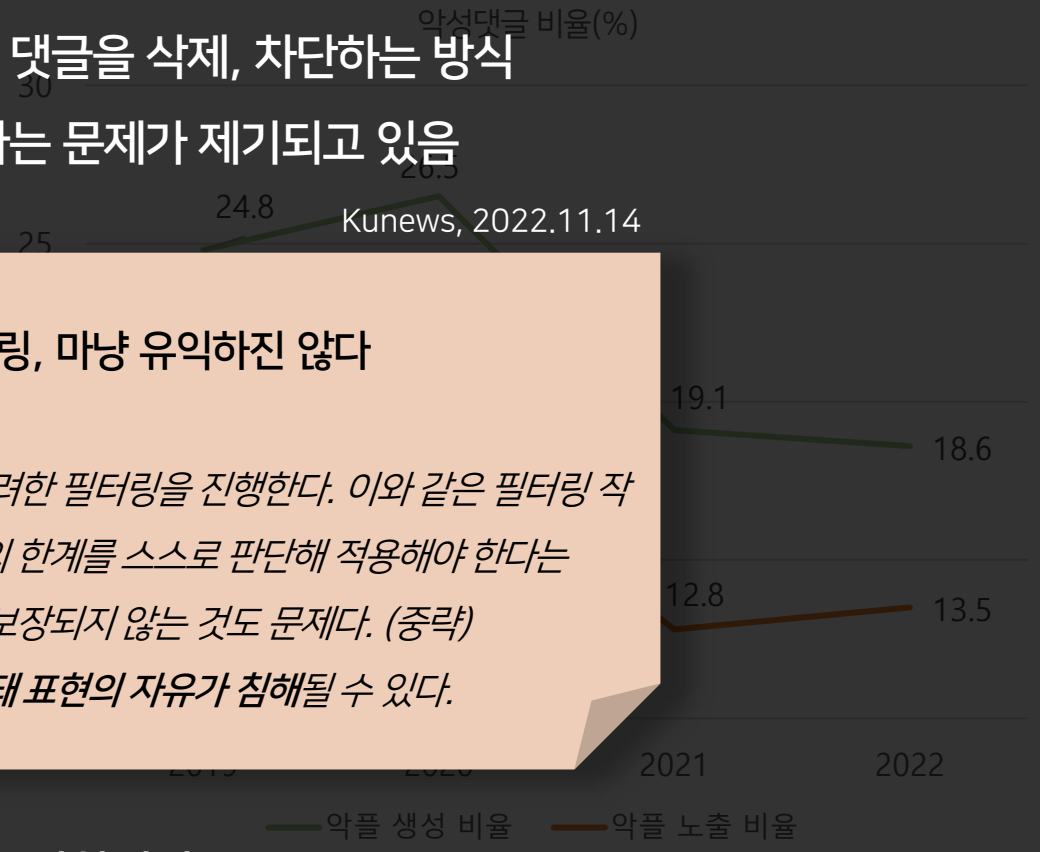
욕설, 비하, 과도한

문장의 전체 맥

'클린봇'의 혐오 표현 필터링, 마냥 유익하진 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.*



표현의 자유를 보장하면서도

AI클린봇 도입을 통해 **악플 생성, 노출 비율 모두 낮아짐**

작성자의 자발적 의지로 악플을 포기하게 만드는

악성 댓글 예방 서비스를 제공하는 것을 목표로 삼음 !

1. 연구 소개

고소 확률 예측 이유

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측

Q1) 왜 고소 여부가 아닌 확률을 예측하나요?

A1) 악플러들에게 더 효과적으로 경고하고, 경각심을 주기 위함입니다!

뭐래 어차피 귀찮아서 고소 안 해ㅋ

ㄴㅋㅋㅋ븅신 집 가서 잠이나 쳐자라

"해당 댓글은 고소당할 수 있습니다."

1. 연구 소개

고소 확률 예측 이유

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측

Q1) 왜 고소 여부가 아닌 확률을 예측하나요?

A1) 악플러들에게 더 효과적으로 경고하고, 경각심을 주기 위함입니다!

ㄴㅋㅋㅋ븅신 집 가서 잠이나 쳐자라

"해당 댓글은 고소당할 확률이 78%입니다."

78%...? 드러워서 안 써

1. 연구 소개

연구 목적

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측

① 표현의 자유를 보장하는 사회문제 해결방법 고안

악성 댓글 삭제가 아닌 고소 확률 제시를 통해 자의적인 악성 댓글 생성 방지

② 건전한 인터넷 문화 형성

악플 작성자뿐만 아니라 일반 대중들에게도 댓글 작성의 영향을 제시해

건전한 인터넷 문화형성에 기여

1. 연구 소개

연구 목적

연구 주제

판례데이터를 활용한 뉴스 댓글 고소 확률 예측

① 표현의 자유를 보장하는 사회문제 해결방법 고안

악성 댓글 삭제가 아닌 고소 확률 제시를 통해 자의적인 악성 댓글 생성 방지

② 건전한 인터넷 문화 형성

악플 작성자뿐만 아니라 일반 대중들에게도 댓글 작성의 영향을 제시해

건전한 인터넷 문화형성에 기여

너무 심한 건 필터링 했습니다.

2. 데이터 수집

댓글 내용 매운맛 수위 알리미

2. 데이터 수집

학습 데이터셋 구성

판례에서 추출한 댓글
(고소label = 1)

고소당하지 않은 일반 댓글
(고소label = 0)

	임베딩벡터	해당연도	…	label
Comment1	…	…	…	1
Comment2	…	…	…	1
Comment3	…	…	…	1
…	…	…	…	1
Comment1710	…	…	…	1
Innocent1	…	…	…	0
Innocent2	…	…	…	0
…	…	…	…	0
…	…	…	…	0
Innocent1710	…	…	…	0

고소 label 클래스의 비율이1:1로 균형을 이루도록 구성!

(클래스 불균형이 존재하면 학습에 악영향)

2. 데이터 수집

학습 데이터셋 구성

고소여부 판단을 위한 임베딩모델 전이학습에서의 파인튜닝을 위해
라벨링(binary)된 댓글들로 학습 데이터셋 구성

이때 고소여부 label 클래스의 비율이 1:1로 균형을 이루도록 구성!
(클래스 불균형이 존재하면 분류 성능에 악영향을 미치기 때문)

2. 데이터 수집

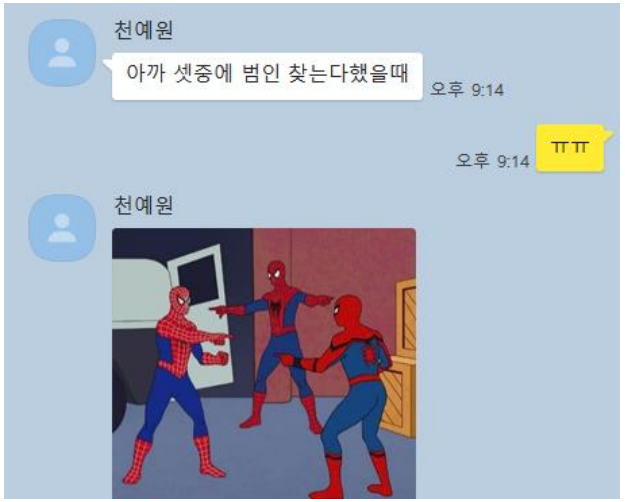
판례데이터 수집 과정 | 국가법령정보센터 API

판례 수집 과정 1

국가법령정보센터 API를 통해

[판시사항, 판결요지, 참조조문, 판례내용] 수집

[누군가가 잘못 다운 받은 상황]



판시사항	판결요지	참조조문	판례내용	고소 label
[1] 모욕죄의 보호법익(=외부적 명예) 및 모욕의 의미 / 표현의 자유와 명예보호 사이의 한계를 설정할 때, …	[1] 모욕죄는 공연히 사람을 모욕하는 경우에 성립하는 범죄로서(형법 제311조), …	[1] 형법 제311조 / [2] 형법 제311조	【주 문】 원심판결을 파기하고, 사건을 서울북부지방법원에 환송한다. 【이 유】 1. 공소사실 요지와 원심 판단: 피고인이 인터넷 포털사이트 뉴스 댓글난에 두 차례에 걸쳐 피해자를 모욕하는 댓글을 게시하였다는 공소사실에 대하여…	1
[1] 어떤 글이 모욕적 표현을 담고 있더라도 사회상규에 위배되지 않는 행위로서 위법성이 조각될 수 있는 경우 …	[1] 모욕죄에서 말하는 모욕이란 사실을 적시하지 아니하고 사람의 사회적 평가를 저하시킬 만한 추상적 판단이나 경멸적 감정을 표현하는 것을 의미한다…	[1] 헌법 제21조, 민법 제750조 / [2] 헌법 제21조, 민법 제750조	【주 문】 원심판결을 파기하고, 사건을 대구지방법원 합의부에 환송한다. 【이 유】 상고이유를 판단한다. 1. 이 사건 공소사실은, 피해자 작성의 “우리에게 ‘독’이 아니라 ‘득’이 되는 MDPS”라는 제목의 기사…	1
…	…	…	…	1

2. 데이터 수집

판례데이터 수집 과정 | 국가법령정보센터 API

국가법령정보센터 API

판례 수집 과정 1

국가법령정보센터 API를 통해

[판시사항, 판결요지, 참조조문, 판례내용] 수집

최종심 판례만 제공해 데이터 부족

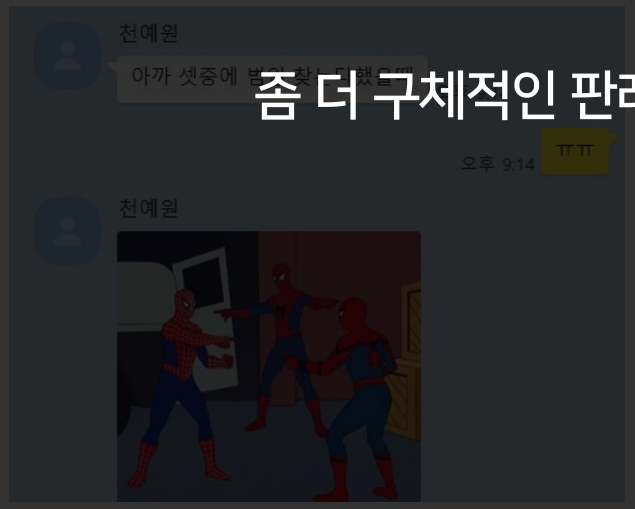
+

판례 내용에 댓글이 존재하지 않는 경우 多

↓

좀 더 구체적인 판례데이터를 제공하는 사이트를 찾아 떠남…

[누군가가 잘못 다운 받은 상황]



판시사항	판결요지	참조조문	판례내용	고소 label
[1] 모욕죄의 보호법익(=외부적 명예) 및 [illegible] 표현의 자유와 명예보호 사이의 한계를 설정할 때, …	[1] 모욕죄는 공연히 사람을 [illegible]에 정립하는 범죄로서(형법 제311조), …	[1] 형법 제311조 / [2] 형법 제311조	【주 문】 원심판결을 파기하고, 사건을 서울북부지방법원에 환송한다. 【이 유】 1. 공소사실 요지와 원심 판단: 피고인이 인터넷 포털사이트 뉴스 댓글난에 두 차례에 걸쳐 피해자를 모욕하는 댓글을 게시하였다는 공소사실에 대하여…	1
[1] 어떤 글이 모욕적 표현을 담고 있더라도 사회상규에 위배되지 않는 [illegible] 성이 조각될 수 있는 경우 …	[1] 모욕죄에서 말하는 모욕이란 사실을 적시하지 아니하고 사람의 사회적 [illegible] 판단이나 경멸적 감정을 표현하는 것을 의미한다…	[1] 헌법 제21조, 민법 제750조 / [illegible] 21조, 민법 제750조	【주 문】 원심판결을 파기하고, 사건을 대구지방법원 합의부에 환송한다. 【이 유】 상고이유를 판단한다. 1. 이 사건 공소사실은, 피해자 [illegible]의 "[illegible]이 아[illegible]" [illegible] …	1
…	…	…		1

2. 데이터 수집

판례데이터 수집 과정 | 로앤비

판례 수집 과정 2

'로앤비'에서 키워드 검색을 통해 댓글이 명시된
판례전문 다운로드 가능 확인!
(다운로드 가능 요금제 월 10만원)

"예로부터 나라의 인재는 성균에 모여 왔으니,
그대 머묾이 우연이겠는가"

명문 성균관과 계약되어 있어 학술정보관을 통해 접속 시 요금제를
이용 가능함을 발견!!

키워드별로 나누어 판례 다운로드

"말로만 듣던 손크롤링과 눈필터링…"

2. 데이터 수집

판례데이터 수집 과정 | 키워드 소개

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

워드클라우드도 참고!

키워드는 사이버 명예훼손, 모욕, 비방 등과 관련된 어휘 또는 법령을 기준으로 설정

사이버 명예훼손죄로 잘 알려짐!

정보통신망이용촉진및정보보호등에관한법률(명예훼손)

제70조(벌칙)

① 사람을 비방할 목적으로 정보통신망을 통하여 공공연하게 사실을 드러내어 다른 사람의 명예를 훼손한 자는 3년 이하의 징역 또는 3천만원 이하의 벌금에 처한다. <개정 2014. 5. 28.>

② 사람을 비방할 목적으로 정보통신망을 통하여 공공연하게 거짓의 사실을 드러내어 다른 사람의 명예를 훼손한 자는 7년 이하의 징역, 10년 이하의 자격정지 또는 5천만원 이하의 벌금에 처한다.

2. 데이터 수집

판례데이터 수집 과정 | 키워드 소개

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

"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).

모욕죄

제311조(모욕) 공연히 사람을 모욕한 자는 1년 이하의 징역이나 금고 또는 200만원 이하의 벌금에 처한다. <개정 1995. 12. 29.>

2. 데이터 수집

판례데이터 수집 과정 | 키워드 소개

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

판례 요약: 롤은 사회악이다…

1. 공소사실: 피고인은 05:30경 불상의 장소에서 '리그 오브 레전드' 게임을 하면서 피해자 B(남, 23세)와 채팅으로 대화하던 중, 피해자에게 "걸레 보지 년임?", "걸레 년인거?", "니 애미 시발년 아니었으면 제드 트롤하기 전에 게임 끝냈다.", "니 애미 시발년이 아무데서나 다리 벌리고 다니니까 이길 게임도 지는거야", "이 제대로 된 새끼한테 질싸 당하고 왔으면 아까 판 이겼어", "걸레 보지년 때문에 이길 사람들이 졌다 하니까", "니 애미 씨발 화냥보지년 때문에 졌다고", "이 씨발 창걸레년의 자식 새끼야", "씹창 걸레년이 방금 판 지게 만들었다", "애새끼 잘못 낳아 기른 뚱 걸레 년", "니 애미 씨발년도 고의로 니 같은 거 낳아서 이긴 판 지게 만듦"이라는 내용의 글을 전송하였다. 이로써 피고인은 자기 또는 다른 사람의 성적 욕망을 유발하거나 만족시킬 목적으로 통신매체를 통해 성적 수치심이나 혐오감을 일으키는 글을 상대방에게 도달하게 하였다.

통신매체이용음란죄

제14조(통신매체이용음란)

자기 또는 다른 사람의 성적 욕망을 유발하거나 만족시킬 목적으로 전화 · 우편 · 컴퓨터 기타 통신매체를 통하여 성적 수치심이나 혐오감을 일으키는 말이나 음향, 글이나 도화, 영상 또는 물건을 상대방에게 도달하게 한 자는 2년 이하의 징역 또는 500만원 이하의 벌금에 처한다.

<개정 2006.10.27>

2. 데이터 수집

판례데이터 수집 과정 | 키워드 소개

키워드	관련 판례 수
정보통신망이용촉진및정보보호 등에관한법률위반(명예훼손)	457
댓글	571
모욕죄	495
비방	3326
통신매체이용음란죄	37
계	4850

한 판례에 여러 키워드가 들어있는 경우 多
이를 중복데이터를 취급하여 제거 후
총 3,541개의 판례 확보

이제 이 판례들에 포함되어 있는 **문제 댓글들을 추출**

아직 멀었다…

2. 데이터 수집

일반 댓글데이터 수집 과정

일반 댓글데이터

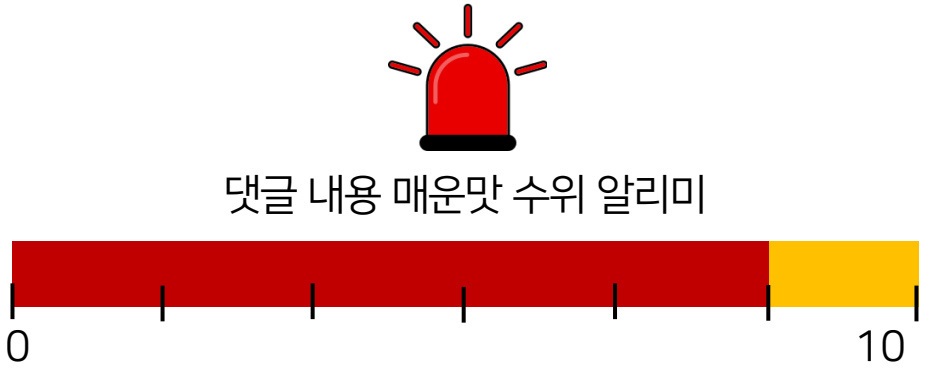
AI HUB에서 **온라인 구어체 말뭉치** 다운로드

- 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) 안 먹으면 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)피디들어오고 나서망했습니다안그러면오래합니다	0
…	…	0

너무 심한 건 필터링 했습니다.

3. 데이터 클렌징



3. 데이터 클렌징

데이터 클렌징

데이터 클렌징

댓글과 발언 위주로 데이터 수집

모델 학습에 사용 가능 여부가 불분명한 데이터에 대해

사용 가능한 기준 제시 후 선별

불분명한 데이터 종류

- 출판물, 게시글
- 다른 죄목
- 법 적용의 다양성
- 문자 메시지
- 상급심 판례

3. 데이터 클렌징

데이터 클렌징

댓글 고소 확률 예측인데 발언도 사용하는 이유

댓글과 발언 위주로 데이터 수집

모델 학습에 사용 가능 여부가 불분명한 데이터에 대해

댓글과 발언의 고소 여부에서 가장 큰 구별점은 공연성(공공성) 성립여부

판례에 나와있는 발언은 이미 공연성이 성립한다는 판단 하에 고소가 이루어진 것

댓글과 발언이 범죄성립요건 중 '공연성'을 공유하기 때문에

판례에서는 비방의 목적이 있는지 여부에 대한 판단

→ 댓글과 발언이 이루어지는 상황을 같은 상황으로 상정

출판물, 게시글 | 다른 죄목 | 법 적용의 다양성 | 문자 메시지 | 상급심 판례

3. 데이터 클렌징

데이터 클렌징

댓글 고소 확률 예측인데 발언도 사용하는 이유

최신판례는 오히려 댓글로 모욕한 경우의 범죄성립요건을 발언의 경우보다 더 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"

이데일리, 2023.11.02

→ 댓글과 발언을 동일선상에 놓고 학습시켰을 때 댓글 고소 확률을 더 정확히 예측할 수 있을 것으로 판단

3. 데이터 클렌징

댓글+발언 판례 수집

댓글 + 발언

비방 목적의 댓글과 발언으로 타인의 명예를 훼손했을 경우 명예훼손죄, 모욕죄 혹은 정보통신망이용촉진및정보보호등에관한법률 위반(명예훼손)죄로 고소당할 수 있음

모욕죄는 사람의 사회적 평가를 저하시킬 만한 추상적 판단이나 경멸적 감정을 표현한 것 혹은 심한 욕설 행위에 적용

정보통신망이용촉진및정보보호등에관한법률 위반죄는 사람을 비방할 목적으로 정보 통신망을 통하여 공공연하게 사실 또는 거짓을 드러내어 다른 사람의 명예를 훼손한 경우에 적용

3. 데이터 클렌징

댓글+발언 판례 수집

댓글 + 발언

비방 목적의 댓글과 발언으로 타인의 명예를 훼손했을 경우 명예훼손죄, 모욕죄 혹은 정보통신망이용촉진및정보보호등에관한법률 위반(명예훼손)죄로 고소당할 수 있음

[댓글 포함 판례 예시]

원고 A는 2018.1.3 피해학생이 자신의 생일축하 댓글에만 답장을 하지 않았다는 이유로 자신의 S에 "나만 차별하는 건가 지랄한다. 짜증나게."라는 글을 게시하였다. 또한 2018.2.14 피해자와 다툰 후 화가 나 자신의 S에 "개짜증나네 ㅅㅃㅂㅆㄷㅍㅂ? ㅂㅅ? ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 하ㅜ", "아 십ㄹㄹ발ㄹㄹㄹㄹ 좇같아요 여러분"이라는 댓글을 게시하였다.

추출한 데이터 : [개짜증나게 ㅅㅂㅅㅂㅆㄷㅍㅂ, 아 십ㄹㄹ발ㄹㄹㄹㄹ 좇같아요 여러분, ……]

3. 데이터 클렌징

댓글+발언 판례 수집

댓글 + 발언

비방 목적의 댓글과 발언으로 타인의 명예를 훼손했을 경우 명예훼손죄, 모욕죄 혹은 정보통신망이용촉진및정보보호등에관한법률 위반(명예훼손)죄로 고소당할 수 있음

피고인은 2012.11.24. 09:40경 대전 00구 00동187에 있는 대전동부경찰서 유치장 장미실에서, 경찰관인 피해자 김00, 한00가 커피 등 편의제공을 해주지 않는다는 이유로 화를 내면서, 위 유치장 장미실 창살을 잡아 흔들며 "지급된 칫솔로 목을 찔러 죽겠다. 상의를 벗어 철창살에 목을 매 죽어버리겠다"라며 소란을 피웠고, 피해자들이 이를 제지하자 피해자 김00에게 "너 개새끼, 니미 씹이다. 니미 씹을 먹고 컸냐, 대가리를 도끼로 쪼개 죽이겠다."라고 욕설을 하고, 피해자 한00에게 "니미 씹할년, 시집도 못가고, 임신도 못하게 보지를 확 찢어 놓겠다."라고 욕설을 하여 다른 수감자들이 있는 상태에서 약 30분 동안 피해자들을 공연히 모욕하였다.

추출한 데이터 : [너 개새끼, 니미 씹이다, 니미 씹할년, ……]

3. 데이터 클렌징

불분명 판례 데이터 필터링

출판물, 게시물 (1)

출판물, 게시물에 관한 판결의 경우, 단순한 미필적 고의에 추가로 주관적 요소인 **비방의 목적** 여부를 중요시하며, 표현 자체에 관한 **제반사정을 감안**하여 훼손된(훼손될 수 있는) **명예의 침해정도**를 고려

…걸쳐 수차 위와 같은 사실 적시를 반복한 점에 비추어 보면, 피고인은 유병을 의도적으로 비방하려 하였다고 볼 것이라고 판단하여, 피고인은 오로지 공공의 이익을 위하여 위와 같은 사실을 적시하였으므로 형법 제310조에 의하여 **위법성이 조각**되어야 한다는 피고인의 주장을 배척하였다. 같은 발언이 위 김춘O 등 6명의 여자들 이외의 불특정 또는 다수인에게 **전파될 가능성이 있다는 점에 관하여는 인식이 없었던 것**으로 봄이 상당하다고 할 것이다.

글에서 명예훼손이 성립되더라도
전파 가능성, 공공의 이익 등
다양한 이유로 명예훼손이 조각될 수 있음

명예훼손이 조각된 출판물 판례는 **제거**

3. 데이터 클렌징

불분명 판례 데이터 필터링

출판물, 게시물 (2)

단, 판례에 **특정 부분을 명시**하여
비방의 목적을 인정해
모욕죄, 명예훼손죄가 성립된 경우에는
판례에 **명시된 부분만** 데이터로 **추출**

(1) 원고 A는 2018.1.3 피해학생이 자신의 생일축하 댓글에만 답장을 하지 않았다는 이유로 자신의 S에 "나만 차별하는 건가 지랄한다. 짜증나게."라는 글을 게시하였다. 또한 2018.2.14 피해자와 다툰 후 화가 나 자신의 S에 "개짜증나네 ㅅㅃㅂㅆㄷㅍㅂ? ㅂㅅ? ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ하ㅜ", "아 십ㄹㄹ발ㄹㄹㄹㄹ 좇같아요 여러분"이라는 댓글을 게시하였다. (중략)

(2) 2018. 2. 24. 피해학생 보호자에게 자신의 잘못을 인정하고 사과하는 문자를 보낸 점 등을 고려하면 위 댓글은 원고 A이 피해학생에 대해 **분노의 감정을 표출함과 동시에 심리적 공격을 가할 목적으로 작성한 것으로 모욕 내지 사이버 따돌림에 준하는 정보통신망을 이용한 언어폭력**에 해당한다.

추출한 데이터 : [나만 차별하는 건가 지랄한다, 짜증나게, ……]

3. 데이터 클렌징

불분명 판례 데이터 필터링

다른 죄목 (1)

판례에 타인을 비방하는 댓글 혹은 발언 등이
존재 하지만 해당 비방이
모욕죄, 명예훼손죄가 아닌
다른 법을 위반한 것으로 고소가 이루어진 경우

제2항을 「2. 같은 해 11. 5. 16:26층 같은 장소에서 위와 같은 인터넷 사이트 게시판에 '친일파에다 가김두한을 사주해 왔던 인물'이라는 제목 하에 "한국역사는 친일과 부패로 얼룩져 있습니다. 아직도 영맥을 잇고 있는 대표적 정당 한나라당과 친일파 대표적 후손 이회창. 나라가 미치지 않은 이 상 이런 자가 대선후보로 나올 순 없습니다."라는 내용을 게시하여 이회창 **후보를 비방**한 것을 비롯 하여 벌지 범죄일람표 2 중 연번 7, 13, 22, 23, 32, 36, 37, 38, 46, 48, 49의 각 기재와 같이 그 시경 1. 범죄사실에 대한 해당법조 판시 제1항의 각 정 : 각 **공직선거및선거부정방지법 제250조 제2항** 판시 제2항의 각 점 : 각 **공직선거및선거부정방지법 제251조** . 판시 제3항의 각 정 : 각 **공직선거및선거부정방지법 제255조 제2항 제5호, 제93조**

비방을 한 사실이 기록되어 있음에도
모욕죄 혹은 명예훼손이 아닌 다른 법(공직선거법 등)으로 판결

해당 데이터가 비방의 목적을 충족하는지
법리적 판단이 이루어지지 않았으므로 **제거**

3. 데이터 클렌징

불분명 판례 데이터 필터링

다른 죄목 (1) 中 상관모욕죄

상관모욕의 경우 일반 상황에 비해 군이라는 엄격한 **수직적 조직의 특성**을 반영해 심리가 이루어지기 때문에 **비방에 해당하는 요건이 완화**되는 특성을 고려하여 상관모욕죄에 해당하는 판례 제거

상관에게는 이정도로 모욕죄 성립…

[수원지방법원 2018. 7. 9. 선고 2017노4615 판결]

"그 후 피고인은 위 상담실을 나가려 하였으나 피해자가 이를 제지하자 소속대 간부 및 병사 7명이 있는 가운데 소리를 지르며 **'안했다고 하지 않습니까? 아침부터 시비 걸어서 사람 아프게 해놓고 이런 것 쓰라고 하는 거는 완전 시비 거는 것이지 않습니까?'**라고 말하여 공연히 **상관인 피해자를 모욕**하였다."

상관모욕죄

군형법 제64조(상관 모욕 등)

① 상관을 그 면전에서 모욕한 사람은 2년 이하의 징역이나 금고에 처한다.

② 문서, 도화(도화) 또는 우상(우상)을 공시(공시)하거나 연설 또는 그 밖의 공연(공연)한 방법으로 상관을 모욕한 사람은 3년 이하의 징역이나 금고에 처한다.

③ 공연히 사실을 적시하여 상관의 명예를 훼손한 사람은 3년 이하의 징역이나 금고에 처한다.

④ 공연히 거짓 사실을 적시하여 상관의 명예를 훼손한 사람은 5년 이하의 징역이나 금고에 처한다.

3. 데이터 클렌징

불분명 판례 데이터 필터링

다른 죄목 (2)

판례의 비방이 다른 법을 위반했지만
모욕죄 혹은 명예훼손죄까지 적용해
같이 처벌받은 경우

↓

해당 데이터가 비방의 목적을 충족하는지
법리적 판단이 이루어진 경우이므로 **추출**

피고인은 2017. 7.4.경 T의 U라는 프로그램에 출연하여 "G이 대통령이 됐는데 조사를 해보니까 부 정선거를 해 가지고 **국민들을 속인 가짜 대통령**이다. 대통령뿐만 아니라 **청와대가 빨갱이 소굴**이 되어 있고, 빨갱이 운동권들이 다 졸중이 모여 있고, E을 부당하게 탄핵하고 사기 선거로써 정권 을 찬탈한 빨갱이 두목 G, "북괴가 지령을 내렸고 빨갱이 정권이 지금 북과 지령을 그대로 실행하 여 E을 구속한 것이다.' "빨갱이 두목 G", 국민을 속여가지고 부정선거로, 투표할 때는 A형, B형 두 종류의 투표지를 가지고 투표를 했는데 개표할 때 보니까 B형은 없어지고 다 A형으로 바꿔치기가 되어 있더라, 투표함 채로 통째로 바꿔치기 했다는 얘기 아니겠습니까:, G 너의 궁극적인 실체는 K 하고 V하고 같이 벌써 수십 년간 저 북괴의 간접 노릇 해 왔잖아. 이거모르는 국민 누가 있어?"라는 인터뷰를 하였다. 이로써 피고인은 위와 같이 **공연히 허위의 사실을 적시하여 피해자의 명예를 훼손**하였다.

범죄사실에 대한 해당법조: 공직선거법 제250조 제2항(허위사실공표의 정, 포괄하여), 공직선거법 제254조 제2항(사전선거운동의 점, 포괄하여), 각 **형법 제307조 제2항(허위사실 적시 명예훼손의 점)**

추출한 데이터 : [국민들을 속인 가짜 대통령, 청와대가 빨갱이 소굴 ……]

3. 데이터 클렌징

불분명 판례 데이터 필터링

법 적용의 다양성

한 판례에서 복수의 법 위반에 대한
심리가 이루어진 경우
모욕죄, 명예훼손죄에 해당한다고
명시된 부분만 데이터로 추출

모욕을 한 경우라도
다른 죄목인 경우는 제외

3. 피해자 O에 대한 **상관모욕**: 피고인은 2020. 7. 초순에서 중순경 사이에 제2항 기재 장소에서, H, J가 듣는 가운데 상관인 중위 O을 지칭하며 "O이 따먹어야겠다. "꼴린다. 라고 수차례 말하여, 공연히 성적 대상으로 희화화하 는 방법으로 상관인 위 피해자를 모욕했다.

7. **모욕**: 피고인은 2020.4.경부터 같은 해 7.경 사이에 공군교육사령부 B학교 G대대일대에서, 1, J 등 다른 병사들이 보는 가운데 피해자 1에 대하여 **'돼지새끼', 허리병신새끼', "뇌까지 장애 있어?"**라고 수 차례 말함으로써, 공연히 위 피해자를 모욕하였다.

모욕죄로 처벌 받은
데이터만 수집

추출한 데이터 : [돼지새끼, 허리병신새끼, ……]

3. 데이터 클렌징

불분명 판례 데이터 필터링

문자 메시지

문자 메시지와 댓글의 가장 큰 차이점은
공연성 성립여부

개인 메시지의 경우 공연성이 성립하지 않음

하지만 전파가능성이 적용되어 공연성을 인정한
판례의 경우, 댓글+발언과 같은 맥락으로 데이터 추출

원고 G은 2018. 1.~2. 무렵 원고 A, T에게 **"U이 병신년 엄마믿고 다댜", "나대", "진짜 개미친년 아니냐, "걍 만원 얼굴에 던져버려 A아", "U네 아빠 눈이 너무 보고싶다"**라는 R **메시지를 보냈다**. 자치위원회 회의록(감 1) 중 43) 간사 사안보고'의 기재에 따르면 이 사건 처분 중 원고 G에 대한 처분의 원인사실이 앞서 인정한 것처럼 원고 G이 원고 A, T에게 보낸 R **메시지 중 피해학생을 대상으로 한 부분**임을 알 수 있고, 그 내용상 모욕 내지 사이버 따돌림에 준하는 정보통신망을 이용한 언어폭력으로 봄이 타당하다.

추출한 데이터 : [병신년 엄마믿고 다댜, 진짜 개미친냔 아니냐, ……]

3. 데이터 클렌징

불분명 판례 데이터 필터링

상급심 판례

상급심 판례의 경우 상소를 거쳐
여러 번 같은 판례 데이터가 등장하지만
중복처리해 제거하지 않고 사용

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

부산지방법원 2019. 8. 23 선고 2019노721 판결

부산지방법원 2019. 2. 14 선고 2018고단452 판결

부산**지방법원** 2019. 2. 14 선고 2018고단452 판결 [아동복지법위반, 정보통신 망이용 촉진및정보보호등에관한법률위반(명예훼손)]

…태메시지에, '**학교폭력범**'이라는 취지로 언급한 것은 피해자를 지칭하는 것이라 할 것이고, '주먹 그 림'을 게시하는 등의 표현 방식과 게시 기간에 비추어 그 의도도 단순히 일반적인 학교폭력방지 목적이라기보다는 비방의 목적이 있다고 인정할 수 있다. 따라서 이 사건 공소사실을 모두 유죄로 인정할 수 있다.

1심 선고에서 추출한 데이터 : [학교폭력범]

3. 데이터 클렌징

불분명 판례 데이터 필터링

상급심 판례

상급심 판례의 경우 상소를 거쳐
여러 번 같은 판례 데이터가 등장하지만
중복처리해 제거하지 않고 사용

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

부산지방법원 2019. 8. 23 선고 2019노721 판결

부산지방법원 2019. 2. 14 선고 2018고단452 판결

대법원 2020. 5. 28 선고 2019도12750 판결 [아동복지법위반 • 정보통신망이용 촉진 및정보보호등에관한법률위반(명예훼손)1 [공2020하, 1298]

는 그 표현의 기초가 되는 사실관계가 드러나 있지 않다. '**학교폭력범**'이라는 단어는 '학교폭력'이 라는 용어에 '죄지은 사람'을 뜻하는 접미사인 '범(3)'을 덧붙인 것으로서, '학교폭력을 저지른 사 람'을 통칭하는 표현인데, 피고인은 '학교폭력범• 자체를 표현의 대상으로 상았을 뿐 특정인을 '학 하시키기에 충분한 구체적인 사실을 드러내 피해자의 명예를 훼손하였다고 보아 이 사건 공소사실 중 정보통신망법 위반(명예훼손) 부분을 유죄로 판단하였다. 원심판결 중 유죄 부분에는 정보통신…

2심, 3심 선고에서 추출한 데이터 : [학교폭력범]

3. 데이터 클렌징

불분명 판례 데이터 필터링

같은 댓글로 상소가 진행된 결과임에도 별개의 데이터로 취급한 이유

대법원 선고에서
'학교폭력범' 데이터 추출

상급심 판례

상급심 판례의 경우
항소, 상고, 항고와 재항고 과
여러 번 같은 판례 데이터가
중복처리해 제거하지 않

대법원 2020. 5. 28 선고 2019도12750 판결

판결 [아동복지법위반 • 정보통신망이용
손)] [공2020하, 1298]

않다. '학교폭력범'이라는 단어는 '학교폭력'이
(犯)'을 덧붙인 것으로서, '학교폭력을 저지른 사
자체를 표현의 대상으로 삼았을 뿐 특정인을 '학
하시키기에 충분한 구체적인 사실을 드러내 피해자의 명예를 훼손하였다고 보아 이 사건 공소사실
중 정보통신망 위반(명예훼손) 부분을 유죄로 판단하였다. 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

3. 데이터 클렌징

최종 판례댓글 데이터셋

	text	comment
1	고등군사법원 2016.6.30,…	네 여자친구랑 섹스해도 되냐?
2	고등군사법원 2017.1.25,…	이게 청소한 거냐, 이 새끼들이 미쳤네 …
3	고등군사법원 2017.1.25,…	넌 새끼야 왜 똑같이 가르쳤는데 쟤보다 못하냐 …
4	고등군사법원 2017.1.25,…	이 미친놈아 …
5	고등군사법원 2017.4.20,…	내가 언제 욕했어, 새끼야.
…	…	…
1707	2002.6.7 선고 2001,…	벤츠를 타고 다니는 온 조합 이사장 자격이 없다
1708	2002.6.7 선고 2001,…	명예와 권위를 짓밟고 욕되게 한 것이므로 …
1709	서울지방법원 2002.9.3 선고,…	나는 공산당이 싫어요
1710	서울지방법원 남부지원 1994.11.16,…	박규는 어떤 인물인가? 박규는 1990년도에 아파트 …

총 1710행의 댓글 / 발언 데이터 확보

3. 데이터 클렌징

일반댓글 데이터

일반댓글 데이터

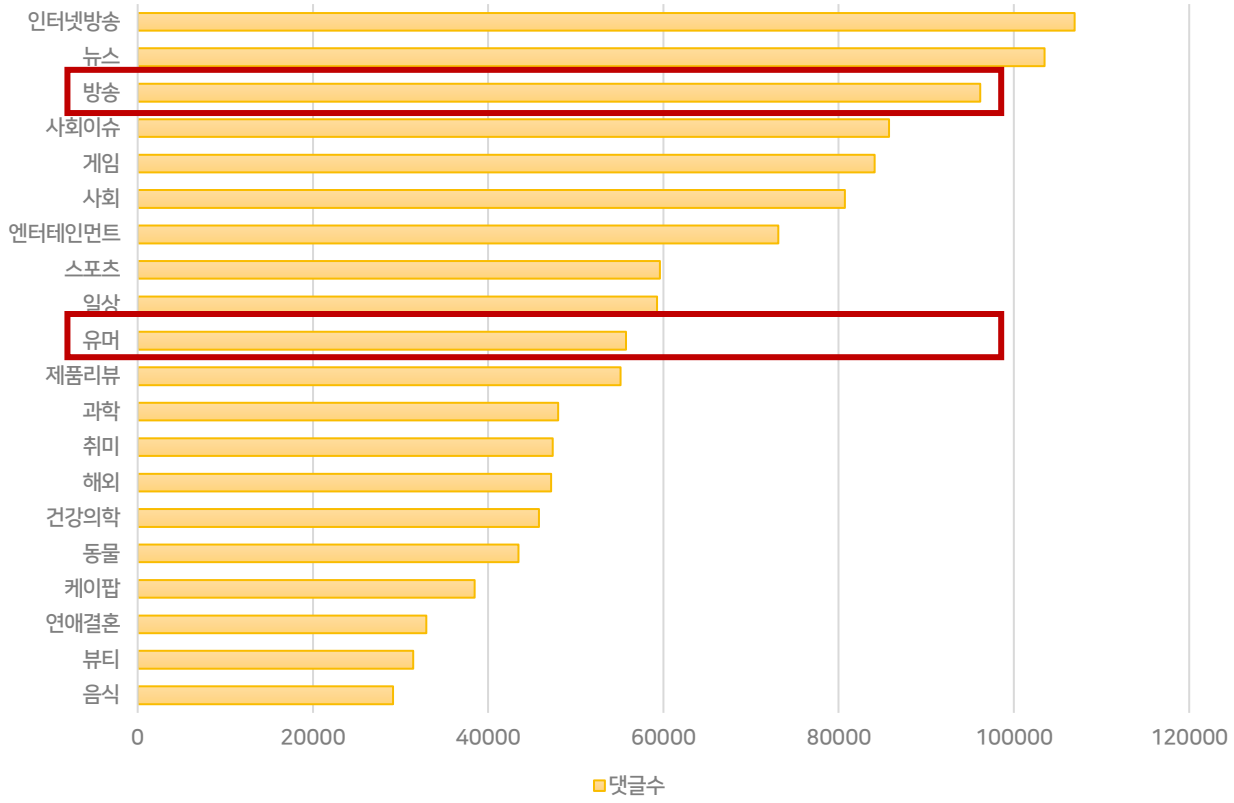
다양한 분야의 댓글 데이터 중

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!

3. 데이터 클렌징

일반댓글 데이터

‘유머’ 분야 댓글

유머 분야 댓글에서 판례와 유사하게
모욕, 비방의 성격을 가진 내용 댓글 다수 발견

⋮

	content
1069	친구 없우세요요유
14171	머리카락이 왤케 없어
14180	진짜 미X놈인줄 알겠다ㅋㅋㅋㅋ

3. 데이터 클렌징

일반댓글 데이터

‘유머’ 분야 댓글
‘유머’ 분야 댓글 475,079행 데이터

	content
0	보물섬분들도 많이 받으세여어어어잇…
1	팝콘 사세요
2	메리 크리스마스.
3	해피 할로윈
4	세복많이 받으세용
…	…
475,076	민석형 더빙너무 잘하거 목소리가 …
475,077	일본편이 잴잼네
475,078	내래이션 은근 잘함 누가함
475,079	썸네일에 (이름) (비속어)가 없다.

3. 데이터 클렌징

일반댓글 데이터

‘방송’ 분야 댓글

방송 분야 댓글에서 판례와 유사하게
모욕, 비방의 성격을 가진 내용 댓글 다수 발견

	content
369	진짜 일머리 겁나없음 혼자 멘붕와서 …
672	강호동이 나이에 비해 생각이 좀 떨어짐
1057	강호동 진짜 제일 답답해 자꾸 행복 저딴 소리나 …

3. 데이터 클렌징

일반댓글 데이터

일반 댓글 데이터 - 방송

'방송' 분야 댓글 584,824행 데이터

	text
0	이수근 너무 화내고 시끄러
1	백번 만 번 돈까스 가격 책정은 이수근이 맞다 봄.
2	여기 왤케 싸우는거임
3	그냥 소스를 조그마한 그릇에 담겨주면 자기들이 …
4	44만원 식재료 판매수익 22만원,
…	
584,821	(이름)이랑 13살(반사회적용어), 12(혐오표현) …
584,822	저거 키스 갑자기 하다 이빨 잘못 …
584,823	여사친 남사친은 실제로 안지 못합니다,
584,824	와시 잠만, 와씨 뭐야 개설레 뭔데 (비속어) …

3. 데이터 클렌징

일반댓글 데이터

이렇게 댓글 유형으로만 구분하고
바로 학습에 사용해도 될까?

주제와 전혀 관련 없는 데이터의 양이 상당해
학습에 방해가 될 것이라고 판단

일반댓글 데이터와 판례 댓글 데이터의
유사도를 계산해 사용 댓글 선정!

자세히는 토큰화 이후에…

3. 데이터 클렌징

임시 일반댓글 데이터셋

일반 댓글 데이터 - 방송
이수근 너무 화내고 시끄러
백번 만 번 돈까스 가격 책정은 이수근이 맞다 봄.
여기 왤케 싸우는거임
그냥 소스를 조그마한 그릇에 담겨주면 자기들이 …
44만원 식재료 판매수익 22만원,
…
(이름)이랑 13살(반사회적용어), 12(혐오표현) …
저거 키스 갑자기 하다 이빨 잘못 …
여사친 남사친은 실제로 안지 못합니다,
와시 잠만, 와씨 뭐야 개설레 뭔데 (비속어) …

일반 댓글 데이터 - 유머
보물섬분들도 많이 받으세여어어잇…
팝콘 사세요
메리 크리스마스.
해피 할로윈
세복많이 받으세용
…
민석형 더빙너무 잘하거 목소리가 …
일본편이 잴잼네ㅋ
내래이션 은근 잘함 누가함
썸네일에 (이름) (비속어)가 없다.

3. 데이터 클렌징

최종 일반댓글 데이터셋(豫)

최종 일반댓글 데이터

일반댓글 데이터의 양이 100만개에 육박하기 때문에 심각한 클래스 불균형 상황

→ 판례 댓글과 일반댓글의 **유사도**를 계산해

판례댓글과 어느정도 **유사한 일반댓글 1710개** 선정 예정

학습데이터 label의 클래스 비율을 1:1로 만들기 위함 !

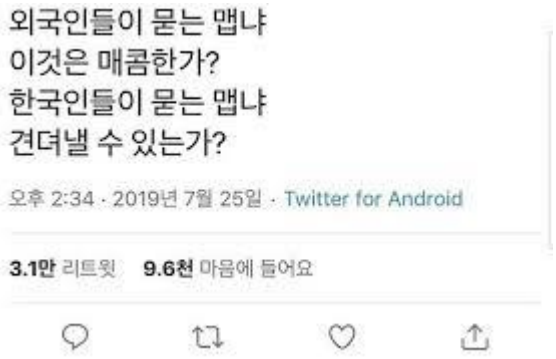
[현재(1주차) 데이터셋]

Label=1인 판례댓글데이터 1710개와

Label = 0인 일반댓글데이터 1710개로 구성

→ 데이터 부족… 돌파구를 찾아보자!

4. 텍스트 증강



안 매워요 = 매콤함

4. 텍스트 증강

데이터 증강

데이터 증강

원본 데이터로부터 **새로운 데이터**를 만드는 기법

학습 데이터 부족으로 과적합 위험이 있거나,

클래스 불균형이 발생했을 때 사용

이미지 증강

- 반전
- 회전
- 밝기 조절
- 블러 처리
- 노이즈 삽입
- 랜덤하게 자르기

텍스트 증강

- SR
- RS
- 역번역
- RI
- RD

4. 텍스트 증강

데이터 증강

데이터 증강

원본 데이터로부터 **새로운 데이터**를 만드는 기법

학습 데이터 부족으로 과적합 위험이 있거나,

클래스 불균형이 발생했을 때 사용

텍스트 데이터는 한 단어만 바뀌어도

문장의 의미가 달라질 수 있기 때문에 주의 필요!

ex) 시계열팀은 패키지를 못 했다 / 시계열팀은 패키지를 안 했다

이미지 증강

- 반전
- 회전
- 밝기 조절
- 블러 처리
- 노이즈 삽입
- 랜덤하게 자르기

텍스트 증강

- SR
- RS
- 역번역
- RI
- RD

4. 텍스트 증강

Back Translation

Back Translation (역번역)

문장을 하나 이상의 다른 언어로 번역한 다음
다시 원래 언어로 번역하는 방법

원문장	번역 언어	번역한 문장	역번역 완료된 문장
씨발, 너 뒤질래 걸레년아	영어	I'm going to look for you, rag	널 찾아갈 거야, 누더기
씨발, 너 뒤질래 걸레년아	프랑스어	Je vais te fouiller.	내가 널 뒤져볼게
씨발, 너 뒤질래 걸레년아	일본어	何でもないですよ、雑巾年よ	아무것도 아니에요, 걸레띠여

너무 구려…

딱 봐도 사용 불가 !

4. 텍스트 증강

텍스트 증강 | EDA

EDA(Easy Data Augmentation)

2019년 EMNLP에서 발표된 논문

"EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks"

해당 논문에서는 4가지의 유용한 텍스트 증강 기법을 소개하고 있음 !

SR

RS

RI

RD

4. 텍스트 증강

텍스트 증강 | EDA

SR	특정 단어를 유의어로 교체하는 방법
RS	임의의 두 단어의 위치를 교체하는 방법
RI	임의의 다른 단어를 삽입하는 방법
RD	임의의 단어를 삭제하는 방법

논문에서는 노이즈를 제공하는 방식을 통해
50%의 train data + EDA로 학습한 결과
100%의 train data와 같은 성능을 확인

한국어로 구현 가능한 코드를 통해 텍스트 증강 진행 !

4. 텍스트 증강

텍스트 증강 | EDA

입력 문장	이 미친놈아 너 나 엿먹이려고 작정했냐
RI + RD	이 미친놈아 너 나 지금 작정했냐
RI	이 미친놈아 너 나 지금 엿먹이려고 작정했냐
RI + RS	이 나 너 미친놈아 지금 엿먹이려고 작정했냐

다른 방법에 비해 비교적 나은 결과를 얻을 수 있었지만,
앞으로 진행될 유사도 계산, 자연어 처리 모델들의 특성상
형태소의 순서가 유의미하다고 판단해 사용하지 않기로 결정

4. 텍스트 증강

텍스트 증강 | EDA

입력 문장	이 미친놈아 너 나 엿먹이려고 작정했냐
RI + RD	이 미친놈아 너 나 지금 작정했냐
RI	이 미친놈아 너 나 지금 엿먹이려고 작정했냐
RI + RS	이 나 너 미친놈아 지금 엿먹이려고 작정했냐

다른 방법에 비해 비교적 나은 결과를 얻을 수 있었지만,
앞으로 진행될 유사도 계산, 자연어 처리 모델들의 특성상
형태소의 순서가 유의미하다고 판단해 사용하지 않기로 결정

이 외에도 임베딩 증강기, CheckList, CLAREA augmentor 등을 사용하였지만,
최종적으로 증강 없이 데이터셋 완성함

5. 자연어 전처리

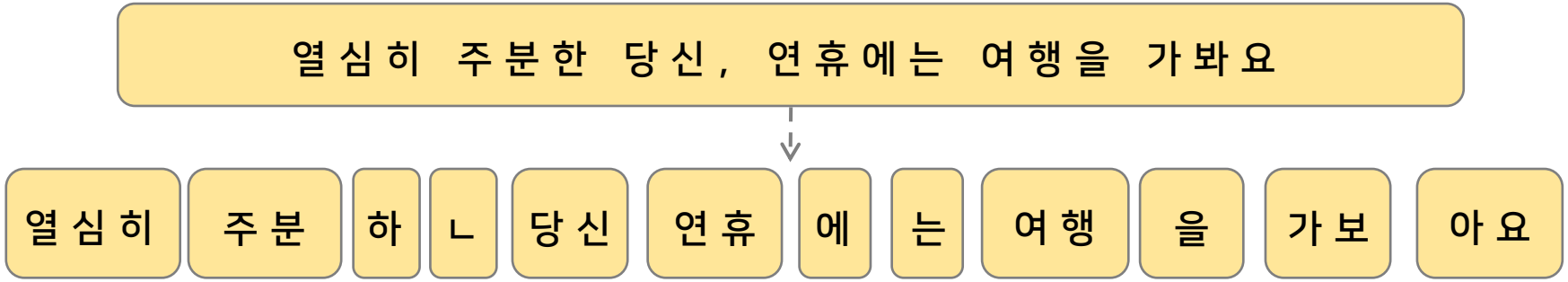
5. 자연어 전처리

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!

5. 자연어 전처리

한국어 토큰화

영어의 경우, 띄어쓰기, 개행, 혹은 구두점 단위로 간편하게 토큰화 할 수 있다는 특징이 있음.

그러나 한국어는 한 띄어쓰기 단위 안에도 여러 개의 형태소가 결합되어 있는 경우 多

희망 가득 찬 여기 쓸데없는 근심 모두 던져버려!

쓸 + 데 + 없 + 는

이러한 한국어의 특징을 살려 토큰화 하기 위해 …

어절 단위 토큰화

형태소 단위 토큰화

Subword 기반 토큰화

음절 단위 토큰화

5. 자연어 전처리

토큰화 단위별 알고리즘

형태소 단위 토큰화

- 꼬꼬마 (Kkma)
- 한나눔 (Hannanum)
- 코모란 (Komoran)
- Okt
- Mecab

Subword 기반 토큰화

- Kiwi
- HuggingFace Tokenizer

형태소 단위로 토큰화 하는 경우 **사전 학습된 형태소 문법** 기반으로 처리가 이루어지므로 문장(혹은 문서) 내 단어의 **등장 횟수와 상관없이 안정적으로 토큰화**가 진행됨.

5. 자연어 전처리

토큰화 단위별 알고리즘

형태소 단위 토큰화

- 꼬꼬마 (Kkma)
- 한나눔 (Hannanum)
- 코모란 (Komoran)
- Okt
- Mecab

Subword 기반 토큰화

- Kiwi
- HuggingFace Tokenizer

형태소 단위로 토큰화 하는 경우 많은 토크나이저가 '**세종 말뭉치**(sejong-corpus)' 기반으로 작동하는데,
만약 토큰화 하려는 문장 속 단어가 해당 말뭉치 사전에 없다면 , 적절한 토큰화가 이루어지지 않음

5. 자연어 전처리

토큰화 단위별 알고리즘

형태소 단위 토큰화

- 꼬꼬마 (Kkma)
- 한나눔 (Hannanum)
- 코모란 (Komoran)
- Okt
- Mecab

Subword 기반 토큰화

- Kiwi
- HuggingFace Tokenizer

Subword 기반 토큰화는 **하나의 단어를 여러 개의 subword로 분리**하여
각 작은 의미의 subword가 문서 내 등장하는 횟수를 토큰화에 반영하므로
희귀한 단어 / 신조어의 토큰화에 효과적 !

OOV에 강인함!

5. 자연어 전처리

토큰화 단위별 알고리즘

우리가 사용할 댓글 데이터의 특성상

형태소 단위 토큰화

- 꼬꼬마 (Kkma)
- 한나눔 (Hannanum)
- 코모란 (Komoran)
- Okt
- Kiwi

Subword 기반 토큰화

- HuggingFace Tokenizer

① 각 어휘들이 가진 미묘한 모욕적 **의미**를 학습하는 것이 중요

형태소 기반 분석기로 의미 중점적 토큰화 必

② 사전 학습되지 못한 **신조어 / 비속어**가 많음

Subword 토크나이저로 신조어/비속어 토큰화 必

Subword 기반 토큰화 : 한 단어를 여러 개의 subword로 분리하여

각 작은 의미의 subword가 문서 내 등장하는 횟수를 토큰화에 반영하므로

희귀한 단어 / 신조어의 토큰화에 효과적 !

따라서, 형태소 토크나이저와 subword 토크나이저를 모두 채택!

OOV에 강인함!

5. 자연어 전처리

토큰화 결과

"씨발, 나이 똥구멍으로 쳐먹지마"

토크나이저	토큰화 결과
한나눔 (Hannanum)	'씨발,나', '똥구멍', '쳐먹지'
꼬꼬마 (Kkma)	'씨', '발', ',', '나이', '똥구멍', '쳐먹', '지', '말'
코모란 (Komoran)	'씨발', ',', '나이', '똥구멍', '으로', '쳐먹', '지', '말', '아'
Okt	'씨발', ',', '나이', '똥구멍', '쳐', '먹지마'
Mecab	'씨발', ',', '나이', '똥구멍', '으로', '쳐먹', '지', '마'
Kiwi	'씨발', '나이', '똥구멍', '치', '먹', '말'

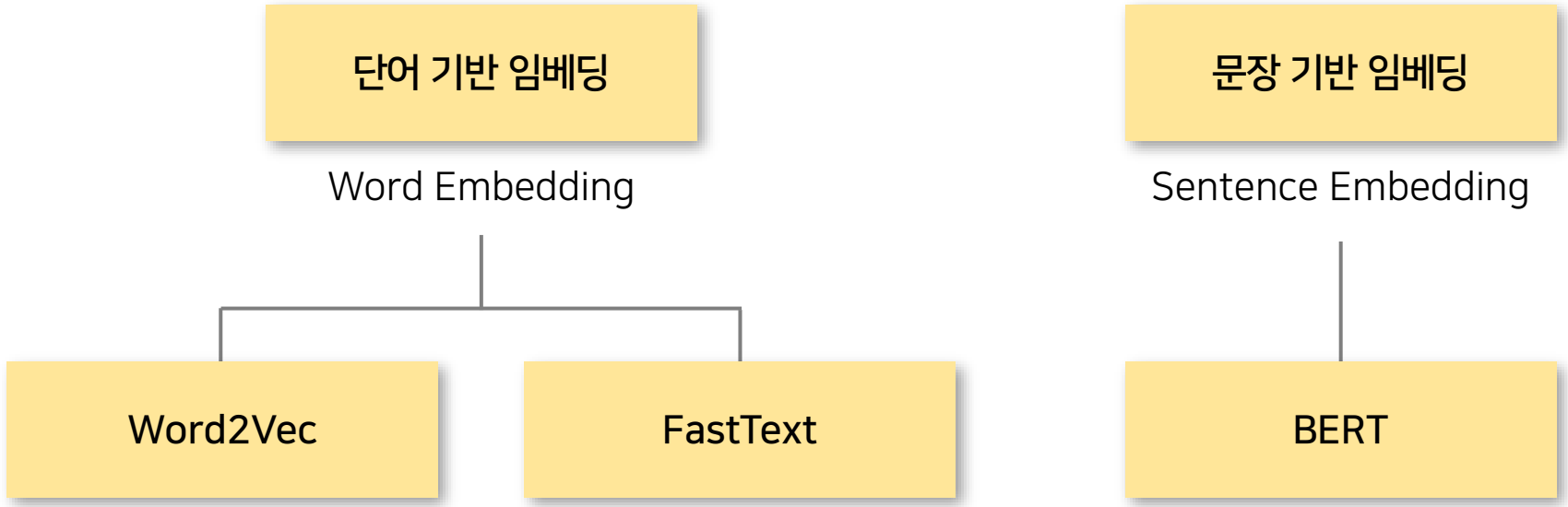
- 한나눔 (Hannanum) : 정제된 언어가 사용되지 않는 경우 형태소 분석 정확도 높지 않음
- Okt : 인터넷 텍스트에 특화된 토크나이저. **비표준어, 속어 등의 처리에 강함**
- Kiwi : Subword 기반 형태소 토크나이저. **비표준어, 속어 등의 처리에 강함**

5. 자연어 전처리

임베딩 (Embedding)

임베딩 (Embedding)

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



5. 자연어 전처리

단어 기반 임베딩 | Word2Vec

CBOW (Continuous Bag of Words)

주변에 있는 단어들을 기반으로 가운데 있는 단어들을 예측하는 방법

중심 단어 주변 단어

Continuous Bag of Words

Continuous Bag of Words

Continuous Bag of Words

Continuous Bag of Words

중심 단어	주변 단어
[1, 0, 0, 0]	[0, 1, 0, 0]
[0, 1, 0, 0]	[1, 0, 0, 0], [0, 0, 1, 0]
[0, 0, 1, 0]	[0, 1, 0, 0], [0, 0, 0, 1]
[0, 0, 0, 1]	[0, 0, 1, 0]

중심 단어 앞 뒤의 주변 단어를 '윈도우' 라고 부름!

5. 자연어 전처리

단어 기반 임베딩 | Word2Vec

CBOW (Continuous Bag of Words)

주변에 있는 단어들을 기반으로 가운데 있는 단어들을 예측하는 방법

이렇게 구해진 원-핫 벡터들은
가중치 W가 곱해진 후
CBOW 모델의 projection layer에서
벡터들의 평균으로 계산

평균 벡터는 두번째 가중치 행렬 W'와 곱해진
후 소프트맥스, 크로스엔트로피 함수를 거쳐
최종 임베딩 형태로 출력

중심 단어	주변 단어
[1, 0, 0, 0]	[0, 1, 0, 0]
[0, 1, 0, 0]	[1, 0, 0, 0], [0, 0, 1, 0]
[0, 0, 1, 0]	[0, 1, 0, 0], [0, 0, 0, 1]
[0, 0, 0, 1]	[0, 0, 1, 0]

5. 자연어 전처리

단어 기반 임베딩 | Word2Vec

Skip-Gram

가운데 있는 단어를 기반으로 주변 단어들을 예측하는 방법

Continuous Bag of Words

Continuous Bag of Words

중심 단어	주변 단어
Bag	Continuous
Bag	Of
Of	Bag
of	Words

전반적으로 Skip-Gram이 CBOW보다 좋은 성능을 가진다고 알려져 있다.

5. 자연어 전처리

단어 기반 임베딩 | Word2Vec

실제 고소 댓글 Word2Vec 임베딩 예시

토큰화 전	“네 여자친구랑 섹스해도 되냐?”
토큰화 후	“여자친구”, “랑”, “섹스”, “해도”, “되냐”
임베딩 후	[-0.4495832 , 0.6018852 , 0.34875992, ……, -0.5376348 , 0.13844709, -0.08259845]

차원 수 조정 가능: (default=100)

5. 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.

"재직증명서" → N-gram (n = 2) → '<재', '재직', '직증', '증명', '명서', '서>'

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

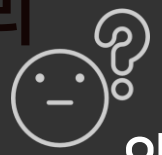
5. 자연어 전처리

단어 기반 임베딩 | FastText

[실제 고소 댓글 FastText 임베딩 예시]

토큰화 전	“네 여자친구랑 섹스해도 되냐?”
토큰화 후	“여자친구”, “랑”, “섹스”, “해도”, “되냐”
임베딩 후	[4.9717464e-02, 6.6876328e-01, -4.1991949e-02, ..., -1.3088505e-01, 6.5155052e-02, 1.1082920e-01]

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

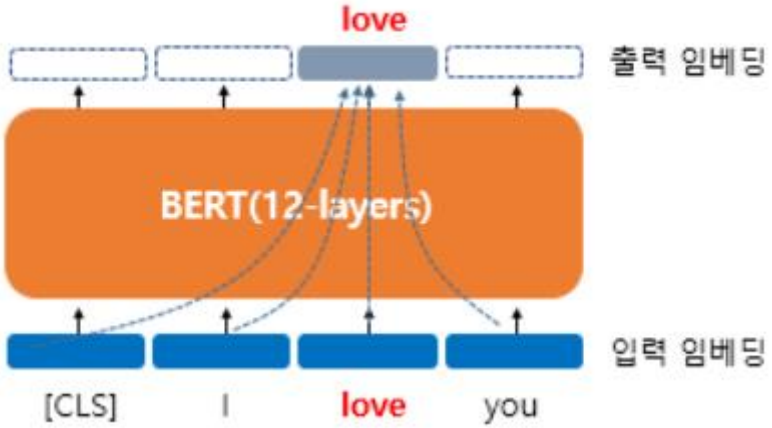
➡ 오타가 많은 댓글(희소 단어로 취급되는 단어)의 처리에 적합함!

5. 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!

5. 자연어 전처리

문장 기반 임베딩 | BERT

[실제 고소 댓글 BERT 임베딩 예시]

원본 댓글	“이 미친놈아 너 나 지금 엿먹일려고 작정했냐, “
임베딩 후	[1.08278580e-01 -1.20382957e-01 2.59359241e-01 -5.17001200e+00 1.92955628e-01 1.31599933e-01 4.25576977e-02 -3.14695872e-02 -2.76154220e-01 -4.25908566e-01 -2.42168352e-01 2.42139488e-01 -1.94791555e-01 -2.95149703e-02 7.04728961e-01 1.11643307e-01 3.58202048e-02 -5.28358556e-02 4.64202277e-02 ….]

5. 자연어 전처리

텍스트 유사도 검사

텍스트 유사도 검사

임베딩된 두 문장(혹은 문서) 사이의 유사도를 측정하는 기법.
유사도를 측정하는 기준으로는 주로 코사인 유사도를 사용하며,
이 밖에 자카드 유사도, 유클리디안 거리 기반 유사도 등이 있음.

Q. 텍스트 유사도를 왜 확인하는데?

A. 실제로 고소당한 판례댓글 데이터(label=1)의 대조군인 일반댓글 데이터(label=0) 중에서 판례댓글과 유사한 댓글 데이터만을 필터링하여 '고소 가능성(likelihood)이 있는 댓글'의 특징을 보다 섬세하게 학습할 수 있도록 하기 위해!

5. 자연어 전처리

텍스트 유사도 검사 | 코사인 유사도, 자카드 유사도

코사인 유사도 (TF-IDF 기반)

텍스트를 TF-IDF으로 벡터화한 후
두 벡터 사이의 각도를
코사인으로 계산하는 방식.

나는 장다연
나는 짱다연
→ 유사도 0.0

형태소 분석 없이 TF-IDF 기반으로
코사인 유사도를 검사하는 경우,
성능이 매우 좋지 않았음.

자카드 유사도

텍스트를 **음절 단위** 집합으로 만든 뒤
두 집합의 **교집합 크기 / 합집합 크기**를
유사도로 사용하는 방식.

구구이즈심심
심구심구즈이
→ 유사도 1.0

음절의 중복만을 평가 기준으로 하기에
문장의 의미를 전~혀 반영하지 못함.
~~여쩌면 잘 작동하는게 이상한 수준…~~

5. 자연어 전처리

텍스트 유사도 검사 | KR-SBERT

SBERT

SBERT는 BERT의 문장 임베딩 성능을 개선시킨(!!!) 모델로,

문장 쌍 회귀 태스크로 모델을 파인 튜닝할 수 있음

이 파인 튜닝 기법을 응용해 문장 유사도 문제(Semantic Textual Similarity) 문제 해결 가능

KR-SBERT는 SBERT모델에 pair꼴을 가지는 KLU-NLI, KorSTS 데이터셋으로 파인튜닝 한 모델

문장 A	문장 B	label
A plane is taking off.	An airplane is taking off.	5.00
A man is playing the piano.	A man seated is playing the piano.	4.25

위와 같이 Label을 유사도로 놓고 문장 A와 문장 B 쌍의 유사도를 학습시킬 수 있음

5. 자연어 전처리

텍스트 유사도 검사 | KR-SBERT 유사도 결과

일반댓글 데이터인 "어느 대학다녀요"와 유사도가 높은 판례댓글 데이터

일반댓글	판례댓글	텍스트 유사도
어느 대학다녀요	학력 컴플렉스	0.385264
	내가 15학번 94년생이고 과에 94년생이 나말고도 세명 더 있다	0.381900
	대머리	0.347509
	아예 그 더러운 놈의 실명을 공개합니다. 저도 명색이 법대출신인데…	0.343811
	…	…
	허리병신새끼	0.000000
	나잇값 쳐먹어라 했다, 새끼야	0.000000

일반댓글 데이터인 "어느 대학다녀요" 문장의 단어 "**대학**"이 가진 의미가 반영되어 …

→ 높은 유사도를 가진 댓글들로 '학력', '학번', '과', '법대' 등 '**대학**'과 관련한 단어가 포함된 문장들이 다수 도출되었음!

3주차 예고

1. 유사도 계산으로 일반데이터셋 필터링
2. 최종데이터셋
3. 추가 변수 선정
4. 모델링
5. 웹 구현

감사합니다