



CNN모델링을 활용한 혐오 댓글 판별기

by 리뷰왕짚천재

이서현

이수빈

전이훈

차은혜



Intro.

- 연구 동기
- 연구 목적



Data.

- 데이터 수집
- 전처리
- 코사인 유사도



Modeling.

- 코드 파일 소개
- 토큰화
- 어텐션 기반 다중채널 CNN
- Overfitting



Result.

- 코드 파일 소개
- 어텐션 기반 다중채널 CNN
- 분석 결과



Concl.

- 결론 및 제언
- 참고 문헌 목록

Intro.





70%

“혐오표현 접한 적 있다”



83%

“온라인에서 혐오 표현을 접했다”



「혐오표현 리포트」

청소년 500명 대상 조사

- 국가인권위원회, 2019.05



워드클라우드

크롤링한 영화 평점 1점 댓글

82년생 김지영



공작



남산의 부장들





사회적 해악

- 인간 존엄과 가치 부정
- 개인적 인권 침해
- 통합 저해



표현 제한의 정당성

- 헌법상 가치인 '민주주의'와 '인권'을 침해



연구 목적

혐오 표현 필터링을 통한 바른 네티즌 문화 형성

- 개별 댓글의 혐오 발언 여부 판별
- 네이버 영화 평점 중 혐오성 댓글을 제외한 실 평점 예측

Data.





DATA

평점

네타즌 관람객 평점

가자-평론가 평점

내 평점 등록하기

한줄평

네타즌 관람객 평점

가자-평론가 평점

내 평점 등록하기

평점

네타즌 관람객 평점

가자-평론가 평점

내 평점 등록하기

한줄평

네타즌 관람객 평점

가자-평론가 평점

내 평점 등록하기

한줄평

총 40,424건

★★★★★ 10

영화내내 가족에게 잘해야겠다 생각이 들었다.우리 아버지는 날려해 모든것을버치신사람이다. 효도해야겠다.

ph08**** | 2014.12.17 00:20 | 신고

👍 9370

🗨 1340

★★★★★ 10

이분들이 있었기에 지금의 내가있고 대한민국이 있는거지. 존경스럽다

imitation of life(days****) | 2014.12.17 01:15 | 신고

👍 8764

🗨 1244

★★★★★ 10

광부, 간호원 여러분! 나라가 부국하고 내가 부국해 여러분이 여기 있습니다! 가족과 고향 생각에 괴로움이 많은 중 알지만 우리가 무엇 때문에 이 먼 이국에 왔는가를 생각해봅시다. 비록 우리 생전엔 이룩하지 못하더라도 후손들에게 잘 사는 나라를 물려줍시다

골레도트(raid****) | 2014.12.17 04:34 | 신고

👍 7529

🗨 1116

네이버 영화 '평점'
데이터 크롤링

A	B	C	D	E	F	G	H
time	ID	text	score	like	dislike	watch_movie	
0 2019-10-23 10:1	점점(jdp****)	본 사람만 평점		10	31682	21409	FALSE

A	B	C	D	E	F	G	H
time	ID	text	score	like	dislike	watch_movie	
0 2018-08-08 9:03	리재원(lips****)	갑자기 분노를		10	5114	2326	FALSE

A	B	C	D	E	F	G
time	ID	text	score	like	dislike	watch_movie
2014-12-17 0:20	ph08****	영화내내 가족에		10	9370	1340 FALSE
2014-12-17 1:15	imitation of life(d	이분들이 있었거		10	8764	1244 FALSE
2014-12-17 4:34	글래도르(raid****	광부, 간호원 여		10	7529	1116 FALSE
2014-12-17 3:57	yon****	주논산 한파를		10	5961	957 FALSE
2014-12-17 11:2	kill****	20대초반 젊은이		8	4788	778 TRUE
2014-12-17 13:5	클로버(ej2r****)	아 눈물범벅. 고		10	3884	366 TRUE
2014-12-17 13:0	조성호(wh48****)	조조로 보고 왔		10	3545	586 TRUE
2014-12-17 0:54	예쁜고기만두(vn	내일을 조국과		10	3725	787 FALSE
2014-12-17 0:32	마땅한게없네(1C	시사회때 봤는데		10	4141	1494 FALSE
2014-12-17 14:4	나두아간다(kjyz*	황정민이캐릭터		10	2405	224 TRUE
2014-12-17 0:38	dydw****	감정 개입 하지		9	2132	488 FALSE
2014-12-17 20:0	손톱달(yang****)	진짜완전대박이		10	1608	181 TRUE
2014-12-17 4:27	취접냉혈(hyun**	내 아버지가 그		10	1651	391 FALSE
2014-12-17 13:0	웅크(xezo****)	울고 웃고 감동		10	1394	154 TRUE
2014-12-17 20:4	애니타임(cont**	정말 우리는 그		10	1326	109 FALSE
2014-12-17 18:4	cool****	전술과 여론이		10	1151	150 TRUE
2014-12-17 20:5	daw****	별점데려하는 자		10	1154	216 FALSE
2014-12-17 12:1	사륜스댕(chon**	가족을 먹여살리		10	1023	111 FALSE
2014-12-17 12:1	dccb****	재밌네요 재밌어		10	1278	409 TRUE

Train data
82년생 김지영
공작
남산의 부장들
국제시장

Test data

캡틴마블
공조
택시운전사



참고: 영화 설명



〈82년생 김지영〉(2019)

- 80년대생 여성의 삶을 다룬 동명 소설을 원작으로 하는 영화로, 성별 갈등을 조장한다는 논란이 일었던 영화이다.



〈공작〉(2018)

- 정보사 소속 스파이가 북한 고위층에 잠입하는 영화. 북한의 대남 도발을 정당화하는 내용으로 논란이 있었다.



〈남산의 부장들〉(2020)

- 1979년 박정희 대통령 암살 전의 서술하는 소설 기반 영화. 대통령 암살범을 미화하고 박정희 대통령을 폄하하였다고 보는 비판적 시선이 존재하는 작품이다.



〈국제시장〉(2014)

- 대한민국 현대사를 두루 겪어온 인물의 일대기를 담아낸 작품. 유신 정권의 행보를 미화하는 요소가 있다는 정치적인 시선이 존재한다.



참고: 영화 설명



〈캡틴 마블〉(2019)

- 마블 유니버스의 히로인 ‘캡틴 마블’주연의 영화. 주연 배우의 페미니즘 및 정치성 발언으로 논란을 겪었다.



〈공조〉(2017)

- 남북한 형사가 공조 수사를 벌이는 영화. 소재와 주연 배우로 인해 북한을 미화하였다는 논란이 있다.



〈택시 운전사〉(2017)

- 1980년 광주 민주화운동을 취재하러 가는 기자를 태우고 광주에 간 택시 운전사의 이야기. 현대사의 민감한 사건을 묘사함에 있어 비판의 시선이 존재한다.



TRAIN DATA

A	B	C	D	E	F	G	H
time	ID	text	score	like	dislike	watch_movie	
0 2019-10-23 10:1	점점(jdp****)	본 사람만 평점	10	31682	21409	FALSE	

A	B	C	D	E	F	G	H
time	ID	text	score	like	dislike	watch_movie	
0 2018-08-08 9:03	리재원(lips****)	갑자기 분노를	10	5114	2326	FALSE	

A	B	C	D	E	F	G
time	ID	text	score	like	dislike	watch_movie
2014-12-17 0:20	ph08****	영화내내 가족에	10	9370	1340	FALSE
2014-12-17 1:15	imitation of life(d	이분들이 있었기	10	8764	1244	FALSE
2014-12-17 4:34	글레도르(raid****	광부, 간호원 여	10	7529	1116	FALSE
2014-12-17 3:57	yon****	추운날 한파를 뚫	10	5961	957	FALSE
2014-12-17 11:2	kil****	20대초반 젊은이	8	4788	778	TRUE
2014-12-17 13:5	클로버(ej2r****)	아 눈물범벅, 고	10	3884	366	TRUE
2014-12-17 13:0	조성호(wh48****	조조로 보고 왔습	10	3545	586	TRUE
2014-12-17 0:54	예쁜고기만두(vn	내일을 조국과 민	10	3725	787	FALSE
2014-12-17 0:32	마당한게없네(1c	시사회때 봤는데	10	4141	1494	FALSE
2014-12-17 14:4	나두야간다(kjyz	황정민이캐릭터	10	2405	224	TRUE
2014-12-17 0:38	dydw****	감정 개입 하지	9	2113	488	FALSE
2014-12-17 20:0	손톱달(yang****)	진짜완전대박입	10	1608	181	TRUE
2014-12-17 4:27	취접냉월(hyun**	내 아버지가 그	10	1651	391	FALSE
2014-12-17 13:0	몽크(xezo****)	웃고 웃고 감동	10	1394	154	TRUE
2014-12-17 20:4	애니타임(cont****	정말 우리는 그	10	1326	109	FALSE
2014-12-17 18:4	cool****	전율과 여운이	10	1151	150	TRUE
2014-12-17 20:5	dany****	별점테러하는 자	10	1154	216	FALSE
2014-12-17 12:1	샤론스댕(chon**	가족을 먹여살리	10	1023	111	FALSE
2014-12-17 12:1	docb****	재밌네요 재밌어	10	1278	409	TRUE



영화명	time	ID	text	score	like	dislike	watch_movie
국제/사과	2015-01-02 18:45	불꽃벽크(mi79****)	그놈이 그놈이 최만영	1	11	31	FALSE

영화명	time	ID	text	score	like	dislike	watch_movie
국제/남인	2020-01-25 22:05	이리구	이리구 최만영	1	22	45	FALSE

영화명	ID	time	text	score	like	dislike	watch_movie	영화명	
국제/남인	201	201	헝두디(godp****)	2018-08-26 16:05	개스락 최관항	1	6	6	FALSE
국제/남인	201	201	LaLa(wesb****)	2018-08-11 13:44	발견이후 위한 발견이	1	8	10	FALSE
국제/남인	201	201	이경이(eja2****)	2018-08-11 19:52	평점 알아 개쩌네	1	11	5	FALSE
국제/남인	202	202	아구조아(dewy****)	2018-09-15 8:12	최관항적인 정치색	1	4	3	FALSE
국제/남인	202	202	스튜디오(just****)	2018-08-17 1:01	흔한 발견이 영화	1	13	8	FALSE
국제/남인	201	201	동대지기(asse****)	2018-08-11 6:58	최파와 그 정치인	1	12	6	FALSE
국제/남인	201	201	FEBRARY(ajdc****)	2018-08-08 16:50	개노샘입니다	1	7	9	FALSE
국제/남인	201	201	범부Fun(yjfo****)	2018-08-15 8:23	황정민 저가	1	5	8	FALSE
국제/남인	201	201	왕주(redb****)	2018-08-29 15:57	보고니왔는데	1	5	6	FALSE
국제/남인	201	201	형인(inbe****)	2018-08-10 23:49	오바다 진짜로	1	8	5	FALSE
국제/남인	201	201	DOTXPIC(kjh****)	2018-10-20 21:10	정치적 색깔이	1	9	5	FALSE
국제/남인	201	201	비디(dbll****)	2018-08-15 21:25	평점이 1 아니면	1	2	4	FALSE
국제/남인	201	201	cjav****	2019-10-13 13:25	진짜 감동이나	1	7	7	FALSE
국제/남인	201	201	태양처럼완전미	2018-08-09 23:54	평점 보니	1	12	6	FALSE
국제/남인	201	201	이치호(hope****)	2018-08-24 21:30	간절으로 국가보안	1	5	6	FALSE

크롤링한 데이터 중 혐오 표현이 많을 것이라고 예상되는

평점 1점 데이터를 영화별 21%씩 랜덤추출



TRAIN DATA

time	ID	text
2020-12-29 11:49	mrbi****	쓰레기 영화 돈OO하지마세요
2019-10-28 12:45	드러머(yosh****)	이세상의 모든짐은 여자가 혼자 다 짊어지고 산다 남자는 이세상의 모든 혜택을 받는다
2019-10-27 9:40	kaz3****	평점9.6 실화냐
2019-10-23 10:30	범쓰(luxb****)	아니 90년대생 여자들이 이걸보고 왜 공감할해 60 70이면 모를까 니들은 왜 공감할해
2019-10-23 18:06	BlueBird(kjr1****)	와 이런 영화가 허허참
2019-10-24 18:33	로센리엔(kist****)	어우 극혐
2019-10-24 21:57	농악먹자(secr****)	그딴 쓰레기 책으로 영화 만든 것 자체가 한 쪽으로 쏠린 평이 나올 수 밖에 없지 남녀
2020-10-11 3:40	최수원 영구제명(qkrc****)	너무 억지고 여자의 부당함만 너무 강조하는 느낌임
2019-10-25 22:52	링통림(hand****)	무슨 SF영화도 아니고 어느 누구의 인생도 힘든일만 모아서 만들면 이정도 안나오는
2019-10-23 10:51	모라고요(jang****)	10대 시절엔 원조 뽀순이로 온갖 사생활위로 사회 이슈 원조교제로 586에게 몸판 세대
2021-03-17 17:57	my food(sss3****)	지가 같은 남편만남길 그 시대의 모든 남편이 그럴거라는 망상하는 한 사람의 이야기
2019-10-24 12:55	이현우(wgus****)	페미는 여성우월주의 입니다
2019-10-25 17:10	김동현(dall****)	본 사람만 평점쓸수있게 해주세요너무 과장되게 평점 싸지르는 인간들이 많아서
2019-11-07 13:17	Aaa(2468****)	도대체 어떤 부분이 공감할부분이라는건지
2019-10-24 16:04	헤플즈(deva****)	어쩌다 있을법한 개인의 안좋은 경험이 누구에게나 있을법한 일처럼 연출되어 남녀갈
2019-10-28 21:56	kadaja222(badr****)	페미코인 극혐
2019-10-24 9:07	시엘(ciel****)	좀 심하네요 원작 소설이 얼마나 문제가 많았는데
2019-12-03 15:54	박상인(tkdd****)	볼 이유가 없음
2019-11-07 14:09	호빵맨(cup7****)	82년생 김민규

4460 rows * 3 columns

전처리 과정

1. 빈 행 제거 및 인코딩 오류난 특수문자 처리

```
score1_text = score1_text.replace("&#34;", "'")
score1_text = score1_text.replace("&#39;", "'")
score1_text = score1_text.replace("&gt;", ">")
score1_text = score1_text.replace("&lt;", "<")
```

2. 앞서 train set으로 선정한 4가지 영화를 하나의 파일로 병합

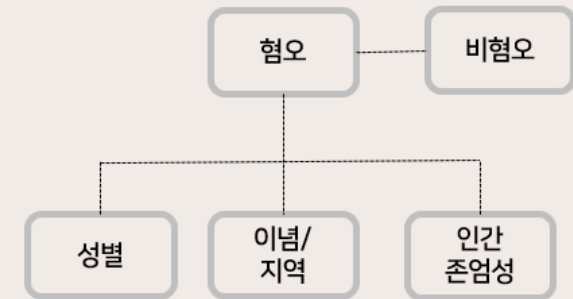
3. 혐오, 성별, 지역/이념, 인간존엄성에 대해 라벨링



TRAIN DATA - LABELING

text	혐오	성별	이념/지역	인간존엄성
쓰레기 영화 돈OO하지마세요	0	0	0	0
이세상의 모든집은 여자가 혼자 다 짊어지고 산다 남자는 이세상의 모든 혜택을 받으며 살아가고있다 82공지영은 아무 실수나 잘못이 없다 0	1	1	0	1
평점9.6 실화냐	0	0	0	0
아니 90년대생 여자들이 이걸보고 왜 공감올해 60 70이면 모를까 니들은 왜 공감올해 머가 와 달았는데	0	0	0	0
와 이런 영화가 허허참	0	0	0	0
어우 극혐	0	0	0	0
그만 쓰레기 책으로 영화 만든 것 자체가 한 쪽으로 쏠린 평이 나올 수 밖에 없지 남녀 불문하고 다같이 힘든데 마치 지들만 힘든일 다 겪은듯이	0	0	0	0
너무 억지고 여자의 부당함만 너무 강조하는 느낌임	0	0	0	0
무슨 SF영화도 아니고 어느 누구의 인생도 힘든일만 모아서 만들면 이정도 안나오는사람 있냐	0	0	0	0
10대 시절엔 원조 빠른이로 온갖 사생행위로 사회 이슈 원조교제로 586에게 몸판 세대20대엔 된장녀라는 신조어를 만들 정도로 골빈 행위를 많	1	1	1	0
지가 같은 남편만난걸 그 시대의 모든 남편이 그럴거라는 망상하는 한 사람의 이야기	0	0	0	0
페미는 여성우월주의입니다	1	1	0	0
본 사람만 평점쓸수있게 해주세요너무 과장되게 평점 싸지르는 인간들이 많아서	1	0	0	1
도대체 어떤 부분이 공감할부분이라는건지	0	0	0	0
어쩌다 있을법한 개인의 안좋은 경험이 누구에게나 있을법한 일처럼 연출되어 남녀갈등을 부추기는 영화	0	0	0	0

4460 rows * 3 columns



각 항목은 중복 라벨이 부여 될 수 있음

혐오적 표현의 경우 전체 train set에서 약 25.4% 차지

혐오적 표현이 아닌 경우 전체 train set에서 약 74.6% 차지



TRAIN DATA – LABELING

YOUTUBE : 증오심 표현에 대한 정책

- 개인이나 집단을 인간 이하로 묘사하거나 동물, 곤충, 해충, 질병 또는 기타 인간이 아닌 다른 대상에 비유하여 존엄성을 침해합니다.
- 앞서 언급된 특성을 근거로 개인이나 집단에 대한 폭력을 찬양하거나 미화합니다.
- 앞서 언급된 특성을 근거로 증오심을 일으키거나 조장하는 인종, 종교 또는 기타 비방이나 고정관념을 사용합니다. 말, 텍스트 또는 이미지의 형태로 고정관념을 조장하거나 사실처럼 전달하는 행위가 포함될 수 있습니다.
- 앞서 언급된 특성을 근거로 개인이나 집단이 신체적 또는 정신적으로 열등하거나 결함이 있거나 병이 있다고 주장합니다. 지능 또는 능력이 모자라거나 장애가 있는 것으로 묘사하여 한 집단을 다른 집단에 비해 열등하다고 표현하는 행위도 여기에 해당합니다.
- 폭력, 차별, 분리 또는 배제를 정당화하기 위해 특정 집단이 앞서 언급된 특성을 지닌 집단보다 우월하다고 주장합니다.
- 앞서 언급된 특성을 근거로 개인이나 집단이 사악하거나 부정적하거나 악의적이라는 음모론을 주장합니다.
- 앞서 언급된 특성을 문제 삼아 개인이나 집단에 대한 예측 또는 지배를 주장합니다.
- 여러 근거를 통해 이미 입증된 폭력적 사건의 발생을 부인합니다.
- 개인이 타인에 대해 갖는 감정적, 낭만적, 성적 호감을 공격합니다.
- 증오를 부추기는 우월주의자들의 신규 활동원 모집이나 자신들의 이데올로기에 대한 재정적 후원 요청이 포함된 콘텐츠입니다.
- 뮤직 비디오의 가사, 메타데이터, 이미지를 통해 증오를 부추기는 우월주의를 조장합니다.

[예제]

코미디영화였으면 10점 실화바탕영화라면 1점 역사를 영화로 배우는 개돼지들

혐오	성별	이념/지역	인간존엄성
1	0	0	1

유튜브 정책 중 ‘증오심 표현에 대한 정책’을 활용하여 전반적인
라벨링 기준 수립



TRAIN DATA - LABELING

애매한 뉘앙스의 댓글 라벨링

“모든 영화는 피해자라는 가정 하에 만들어진 여성혐오 영화”



혐오	성별	이념/지역	인간존엄성
0	0	0	0

혐오 X

“편집증 피해망상 남성혐오 선동 그 자체”



혐오	성별	이념/지역	인간존엄성
1	0	0	1

혐오 O



TRAIN DATA - LABELING

애매한 뉘앙스의 댓글 라벨링

“픽션을 마치 사실인양 만들어 놓은 선동영화”



혐오	성별	이념/지역	인간존엄성
0	0	0	0

혐오 X

“아무리 영화로만 생각하고 보려고 해도 그냥 좌발 선동영화”



혐오	성별	이념/지역	인간존엄성
1	0	1	0

혐오 O



TRAIN DATA - LABELING

데이터 불균형

data[data['성별']=='=0]						
	text	필요	성별	이념/지역	인간존엄성	
0	data[data['성별']=='=1]					
2						
3						
4	1	이세상의 모든것은 여자가 혼자 다 짊어지고 산다 남자는 이세상의 모든 혜택을 받으...	1	1	0	1
5	9	10대 시절엔 원조 박순이로 온갖 사생팔위로 사회 이슈 원조고제로 586에게 물판 ...	1	1	1	0
...	11	페미는 여성우월주의입니다	1	1	0	0
4467	15	페미포인 국립	1	1	0	0
4468	24	페미의 페미에 의한 페미를 위한 영화 따위	1	1	0	0
4469
4470 경제	3230	OOO기 82kg 김지영	1	1	0	0
4471	3236	취미미니를 개국했네요	1	1	0	1
3823 rows	3249	20대 30대 여성들의 수준을 보여주는 영화	1	1	0	0
	3252	애기 얼마입니다 공갈 전하안되고 아메리카노 쏟았다고 말중이라고 욕하는 사람 없습니...	1	1	0	0
	3256	80년대성 여자중에 자발 받으며 커왔고 힘든 시절살아 하는 여자 누가 있는지 진짜...	1	1	0	0
637 rows × 5 columns						

data[data['이념/지역']=='=0]						
	text	필요	성별	이념/지역	인간존엄성	
0	data[data['이념/지역']=='=1]					
1	이세상					
2						
9	10대 시절엔 원조 박순이로 온갖 사생팔위로 사회 이슈 원조고제로 586에게 물판 ...	1	1	1	0	
142	최악의 최악 판가르는 영화가 뭐가 좋나요 다들 형이 페미라서 나라가 이런건가	1	1	1	0	
147	이런 되도록은 보내방상글이 영화로 쳐 올라오는걸 보면 한국은 더 이상 미래가 없다	1	0	1	1	
187	배 처 부른 소리하지들 말고 허잡 쓰고 살아야 하는 나라에서 살아라	1	0	1	0	
...						
4466	248	역겨운 영화 5060년 대성 어머니 세대는 힘들게 사신거 인정한다 이 분들은 페미...	1	1	1	0
4467
4468	4423	일베들이 극찬하는 영화라면서?	1	0	1	0
4469	4427	말에 타성.. 너같은 무지하게 세뇌된 말보수주의자는 이만 쓰러기영화나 전변안변바라 ...	1	0	1	1
4471	4434	0점이 선택이 안된다니 진짜 안타깝네요. 이 영화 좋다는 분들(일베, 일베 말고 있...	1	0	1	0
	4451	완그네가 보고 눈물을 흘렸다고 벼랑이들이 오줌을 지렸다는 그 영화군.	1	0	1	1
4189 rows ×	4470	경제성장이 국면이 최성한 국이라고? 왜 삼성도 노동자가 만들었다고 해보지? 미친 말...	1	0	1	0
271 rows × 5 columns						

data[data['인간존엄성']=='=0]						
	text	필요	성별	이념/지역	인간존엄성	
0	data[data['인간존엄성']=='=1]					
2						
3						
4	1	이세상의 모든것은 여자가 혼자 다 짊어지고 산다 남자는 이세상의 모든 혜택을 받으...	1	1	0	1
5	12	본 사람만 공감할수있게 해주세요너무 과장되게 감정 써지르는 인간들이 많아서	1	0	0	1
	27	물거없이요 이 영화를 보고 형사책이라고 생각한다면 그것 자체가 피해당상임	1	0	0	1
...	33	또시작이네집그러운거	1	0	0	1
4467	57	62년생 김지영이었으면 누구나 공감했을텐남여는 생물학적으로 차이가 있고 그 차이에 ...	1	0	0	1
4468
4469	4394	여이 개돼지새끼들아. 이렇게 우리 부모님들은 개고생하며 몰바쳐 나탈위해 고생하셨다....	1	0	1	1
4470 경제	4422	연기 빠고는 맛이 없을 신과는 국이고 표절도국임이게 재미있다는 대중에 개한스럽다 ...	1	0	0	1
4471	4427	말에 타성.. 너같은 무지하게 세뇌된 말보수주의자는 이만 쓰러기영화나 전변안변바라 ...	1	0	1	1
	4451	완그네가 보고 눈물을 흘렸다고 벼랑이들이 오줌을 지렸다는 그 영화군	1	0	1	1
4033 rows	4454	별별 호텔공식에 어거지 눈물 자아내는 주제를 들고 와도역시 감독이 모자라면 작품이 ...	1	0	0	1
427 rows × 5 columns						

전반적으로 각 항목별 데이터 불균형이 존재하는 경향으로

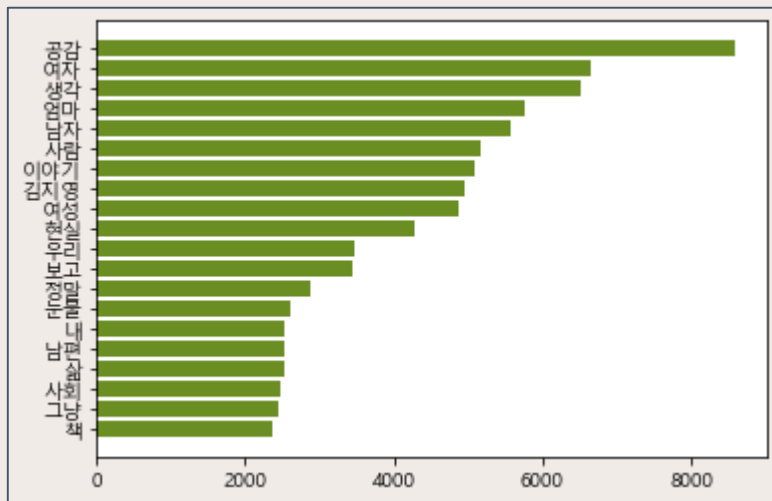
특히 이념/지역의 경우 데이터 불균형이 심한 편



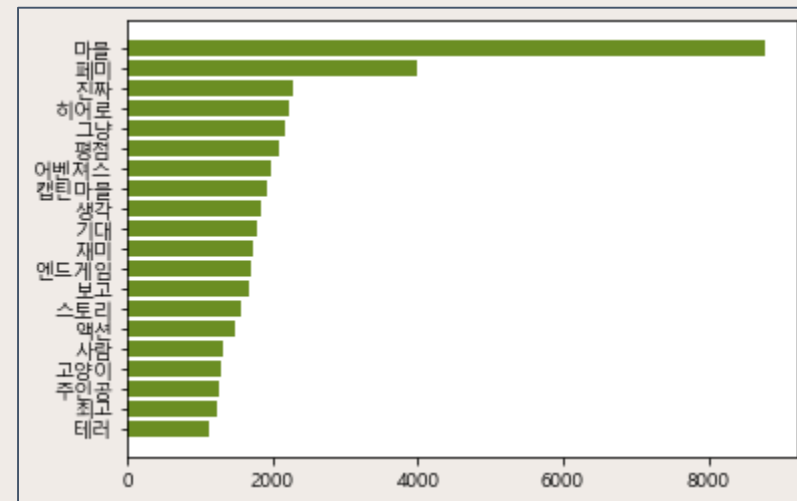
TEST DATA - 선정 방법

코사인 유사도

82년생 김지영



캡틴마블



cosine similarity

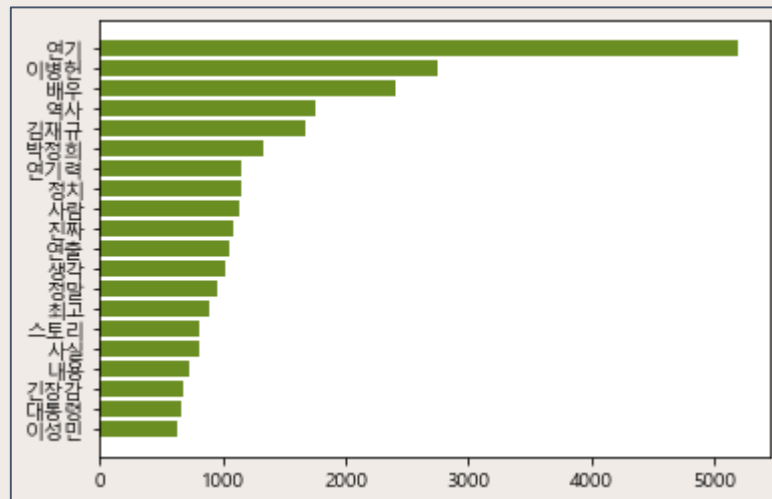
0.508875



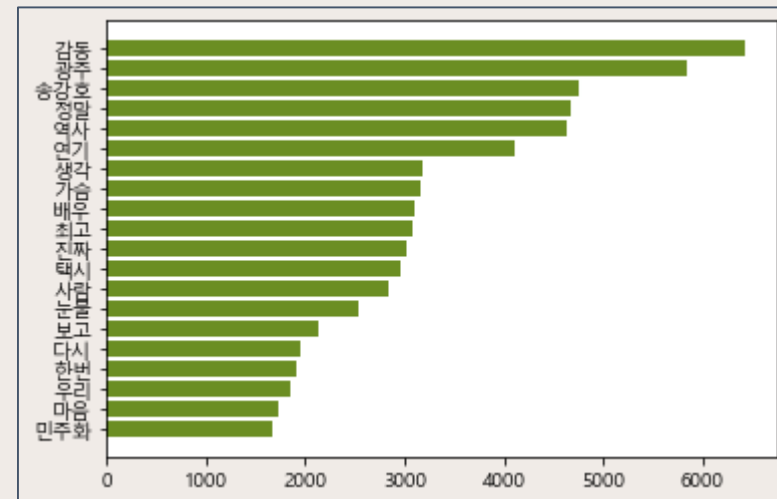
TEST DATA - 선정 방법

코사인 유사도

남산의 부장들



택시운전사



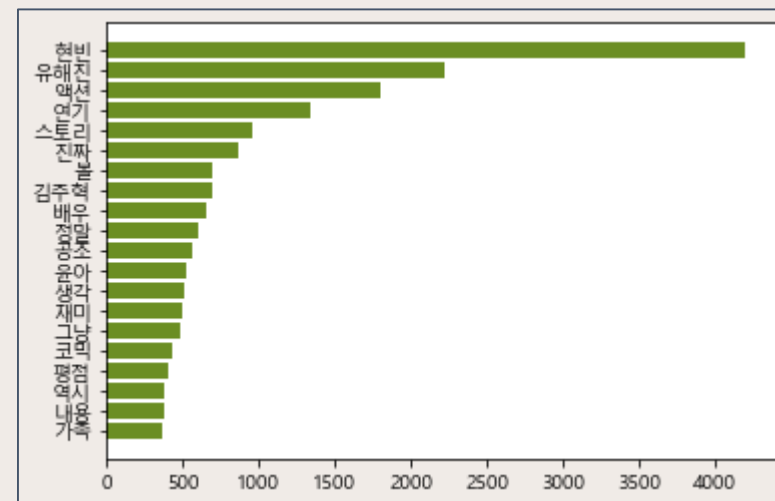
cosine similarity

0.47443452



코사인 유사도

공조



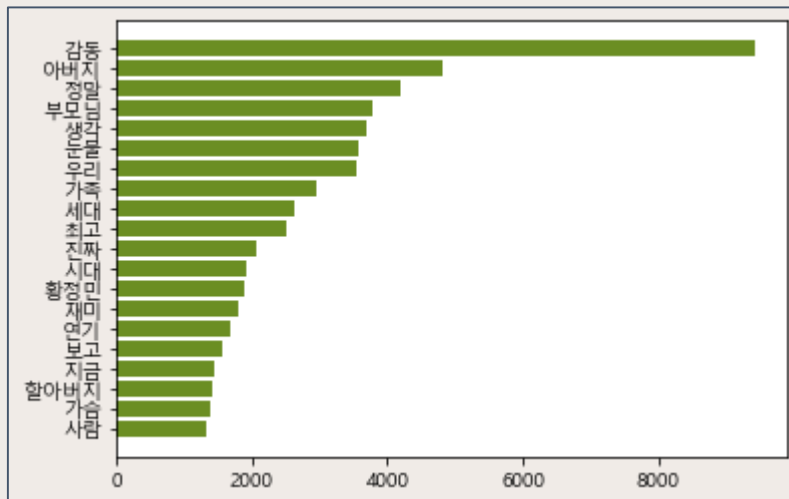
0.66978106



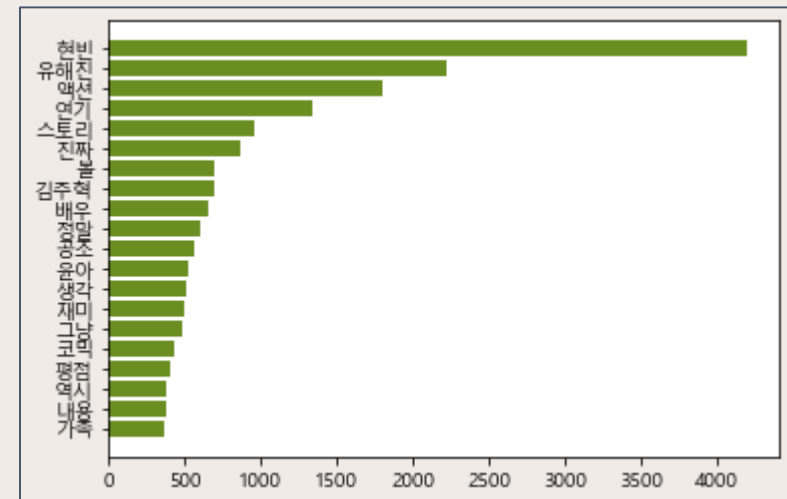
TEST DATA - 선정 방법

코사인 유사도

국제시장



공조



cosine similarity

0.63916115

Model.

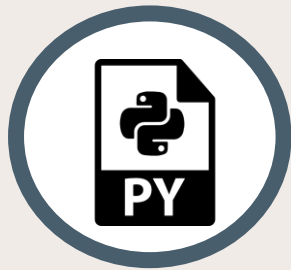




코드파일소개

| 리뷰왕찐천재

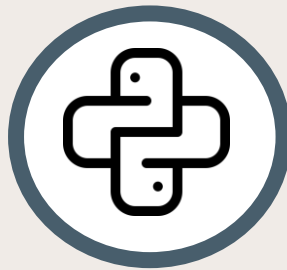
사용된 코드 파일



Token.py

Metric.py

AttentionLayer.py



train.ipynb

test.ipynb



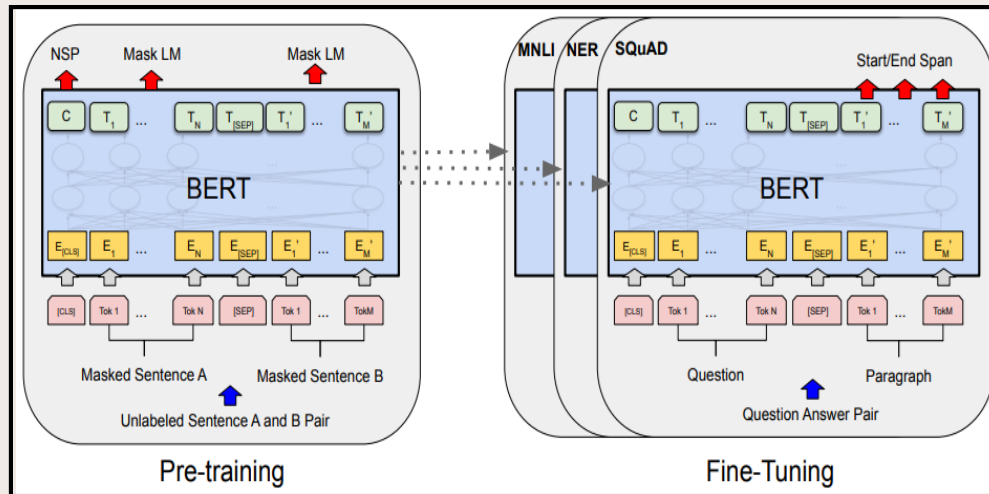
Tokenizer

w2v_pretrain_emb



Huggigface의 transformers 라이브러리

KcBERT

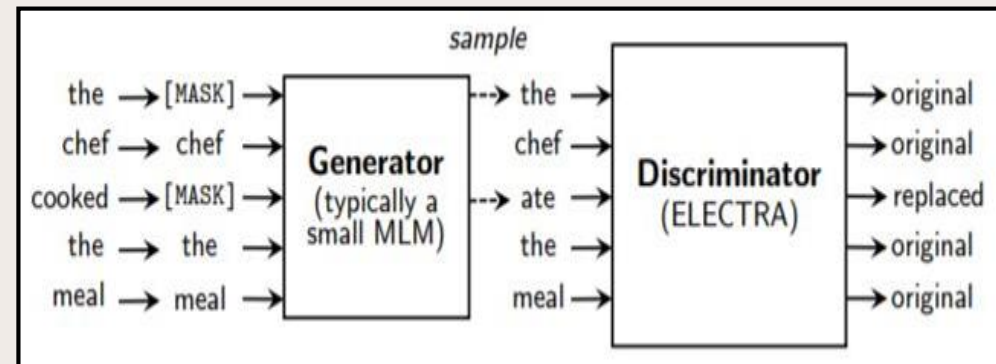


한국어 댓글로 사전훈련된 BERT모델

BERT : Transformer의 Encoder를 여러층으로 쌓아 문맥을 반영한 임베딩(Contextual Embedding)

마스킹 방식의 BertForMaskedLM 모델

KcELECTRA



RTD(Replaced Token Detection) 학습방법

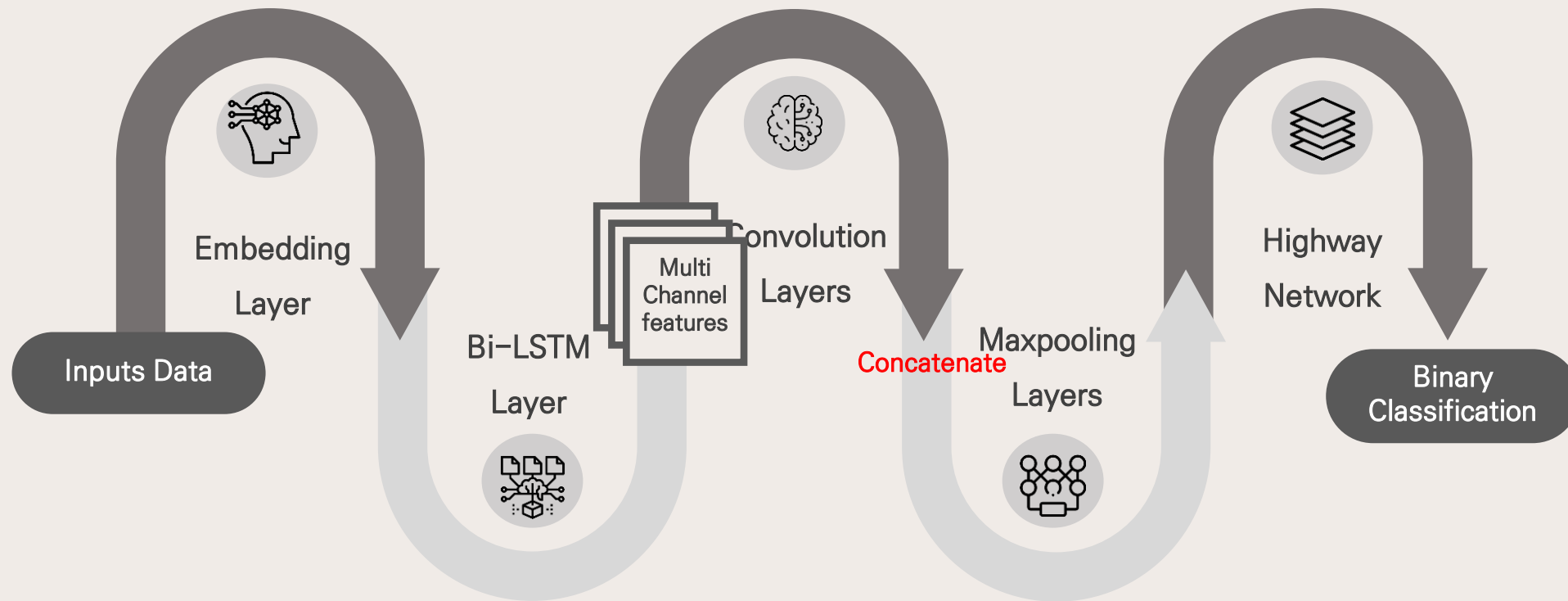
네이버 뉴스에서 댓글수집

KcBERT 대비 데이터셋 증가 및 vocab 확장



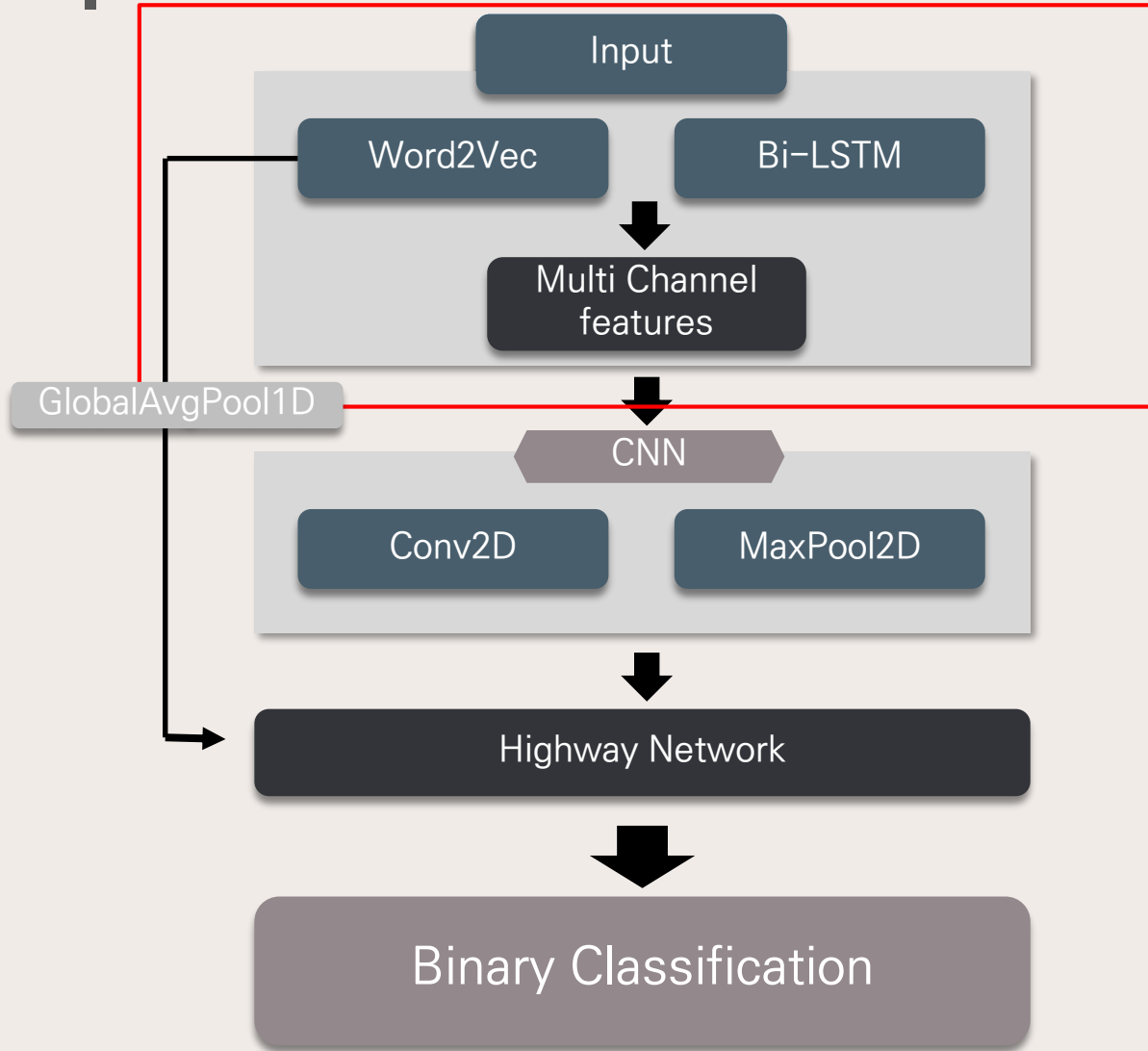
어텐션 기반 다중채널 CNN | 리뷰왕짚천재

Model





Multi-channel CNN Model 기법



Embedding

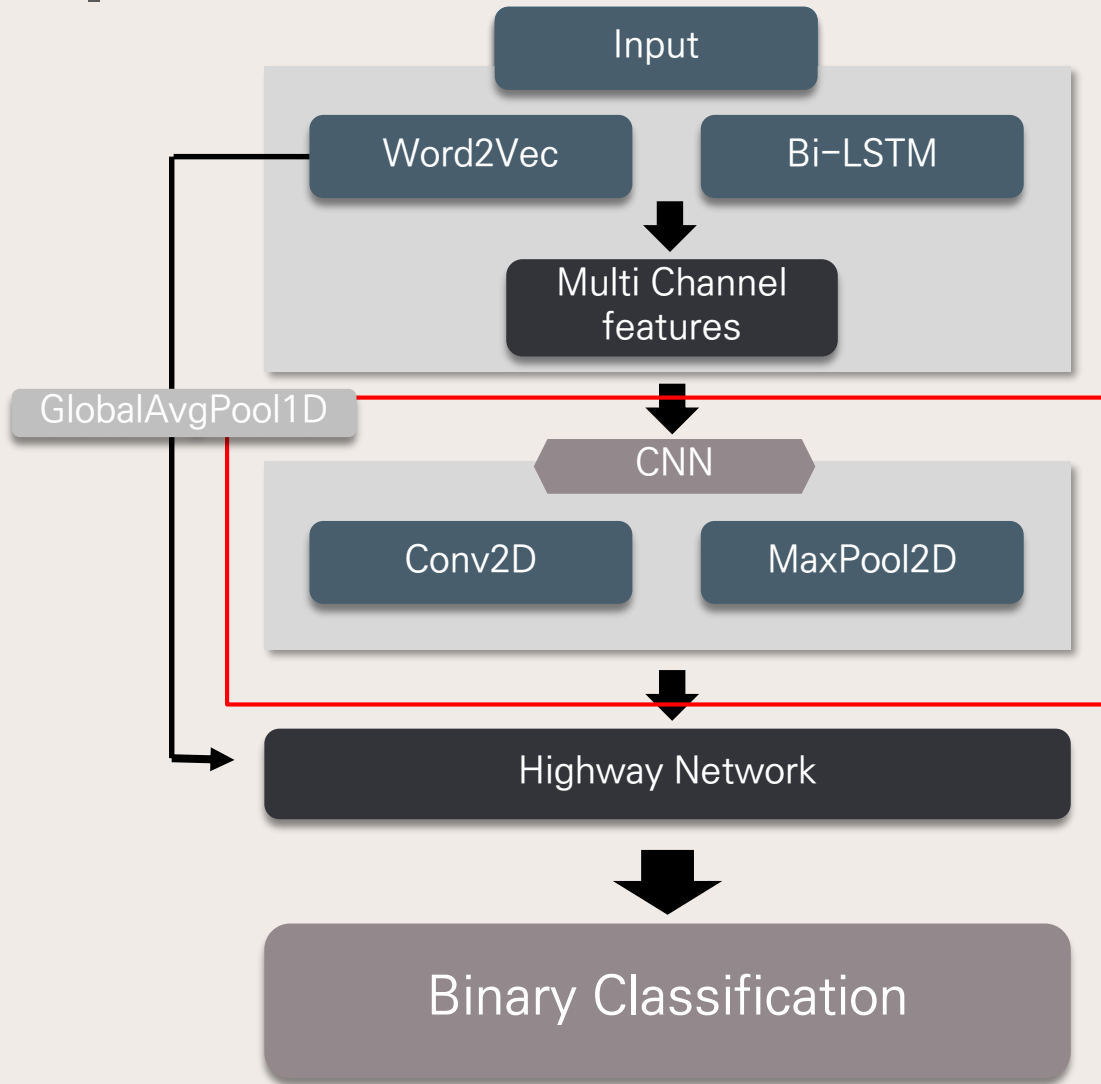
- ➡ Pre-trained word embedding ← Word2Vec
- 단어간 유사도를 반영할 수 있도록 단어의 의미를 벡터화

Bi-LSTM

- ➡ 문장의 양방향에서 문맥에 대한 정보를 추출
- 단어에 대한 양방향 hidden state 얻음
- 어텐션 메커니즘 : 중요도 산출 후 가중치 부여



Multi-channel CNN Model 기법



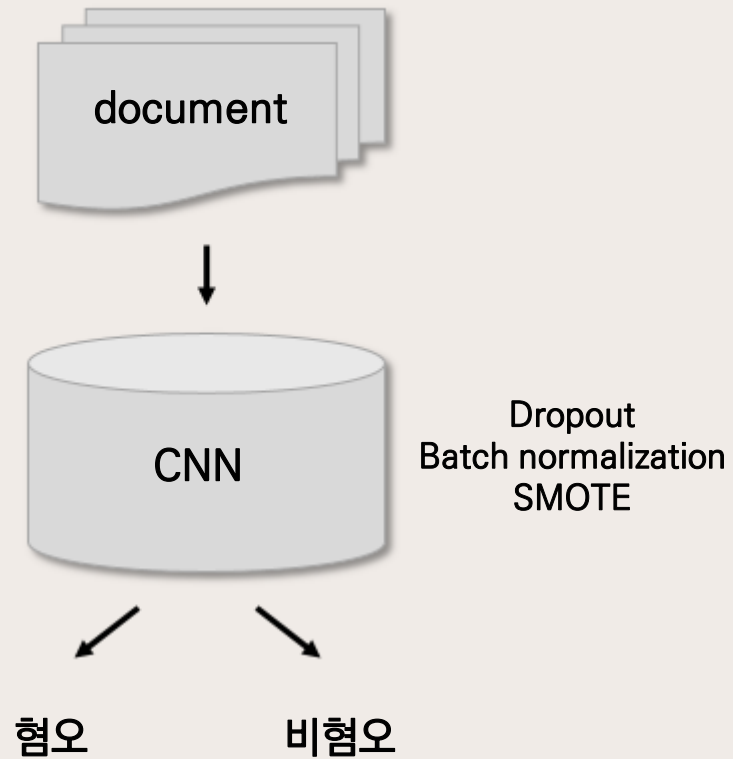
Convolution Neural Network(CNN)

- ➔ 3가지 특성채널이 입력값
- 2D-Convolution & 2D-Maxpooling 연산
- 배치정규화(BatchNormalization) : 과적합 방지
- Dropout(0.6) : 과적합 방지
- 활성화 함수 : sigmoid → relu



과적합 원인

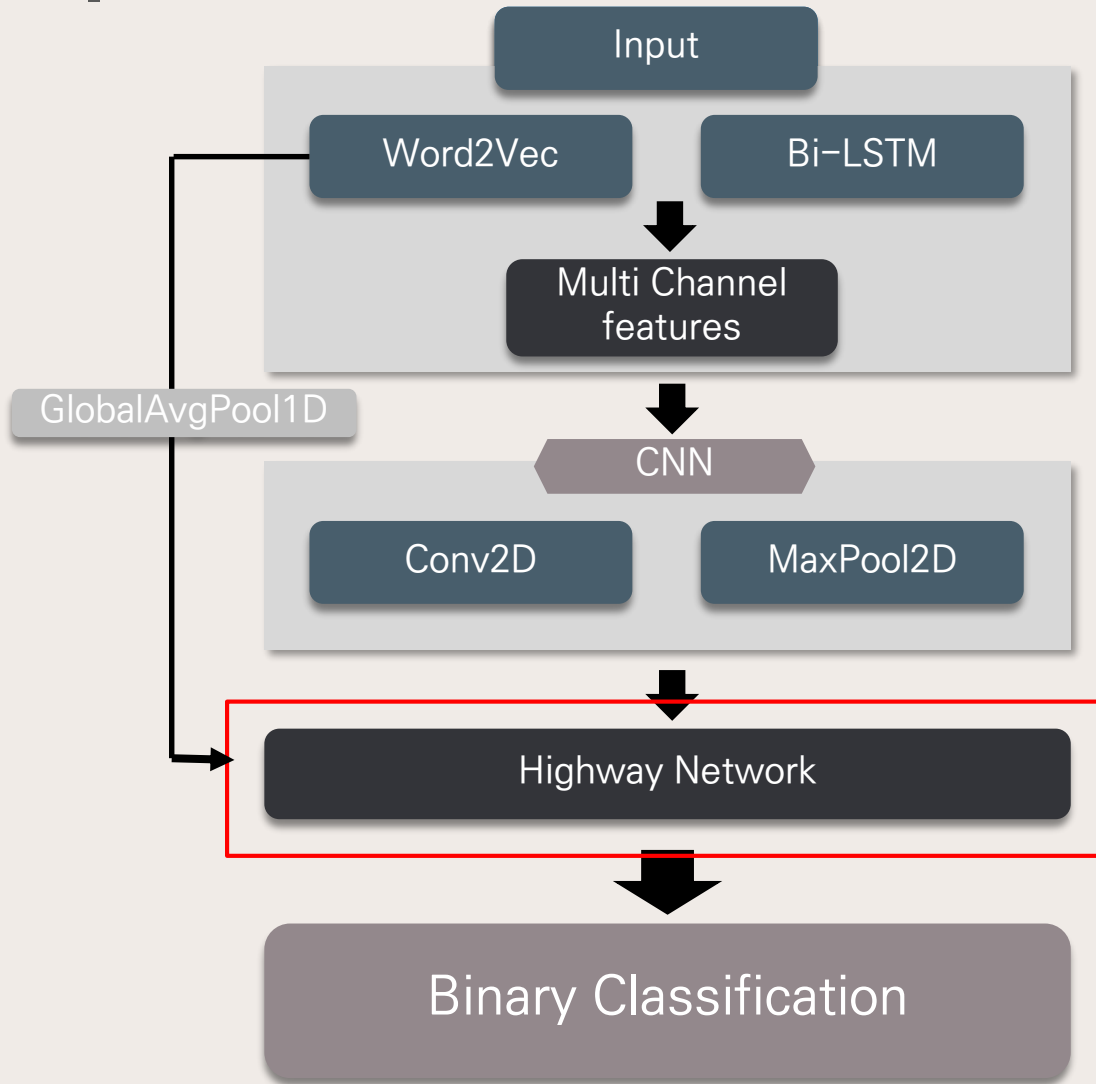
train data의 적은 양 & 클래스 불균형



	시도기법	적용기법
과적합 방지	파라미터 조정 (batch_size , learning rate) 가중치 규제(L2 regularization) K-Fold Dropout Batch normalization	Dropout Batch normalization
불균형 해소	Random Oversampling SMOTE ADASYN(Adaptive Synthetic Sampling)	SMOTE



Multi-channel CNN Model 기법



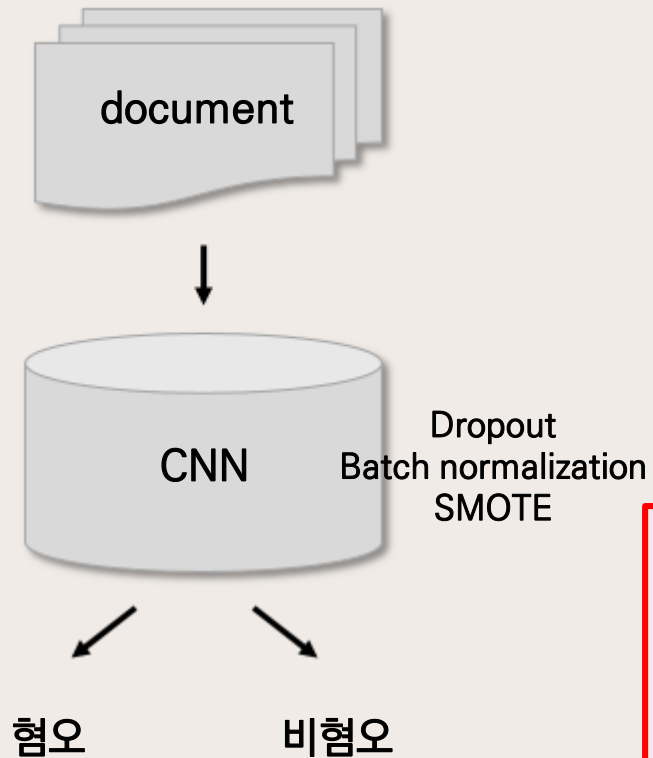
Highway Network

- 복잡한 모델의 정보 흐름을 통제
- 임베딩 출력과 CNN출력의 반영될 정도 결정
- 임베딩 출력값에 1D-Global Average Pooling적용



과적합 원인

train data의 적은 양 & 클래스 불균형



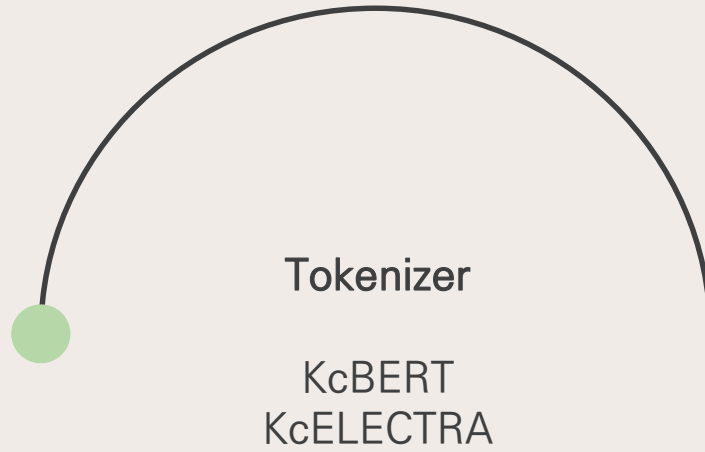
	시도기법	적용기법
과적합 방지	파라미터 조정 (batch_size , learning rate) 가중치 규제(L2 regularization) K-Fold Dropout Batch normalization	Dropout Batch normalization
불균형 해소	Random Oversampling SMOTE ADASYN(Adaptive Synthetic Sampling)	SMOTE

Result.





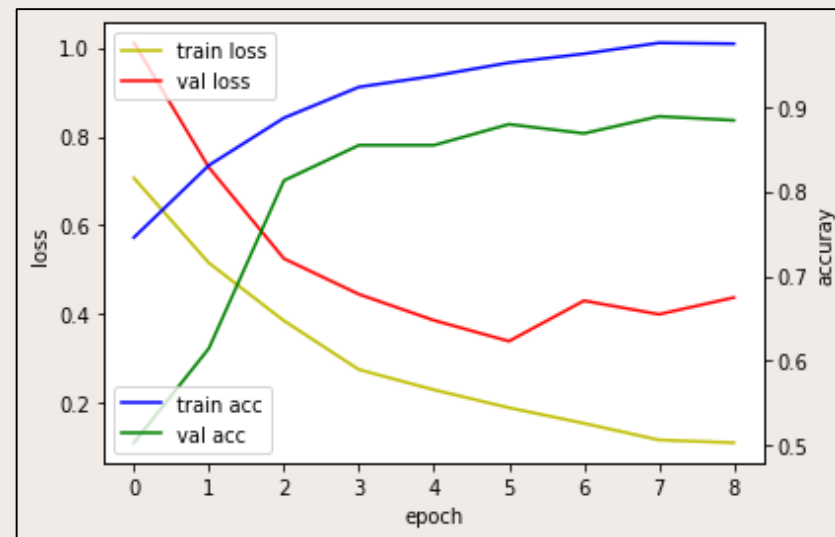
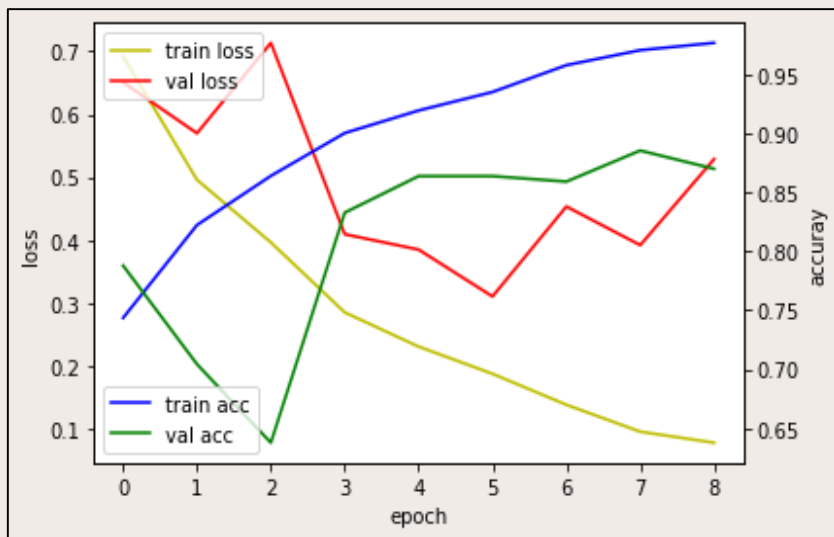
컬럼별 모델 선정 기준





컬럼별 모델 선정 기준

val_loss의 추세



val_loss의 하락세가 강하고, 안정적인 모델을 선택



컬럼별 모델 선정 기준

높은 F1-score

Tokenizer	KcELECTRA		KcBERT	
lr_rate	0.001	0.0005	0.001	0.0005
혐오	0.685	0.690	0.605	0.602
성별	0.570	0.639	0.660	0.502
이념/지역	0.303	0.346	0.342	0.359
인간존엄성	0.220	0.300	0.345	0.373

Tokenizer, 학습률에 따른 F1-score 변화



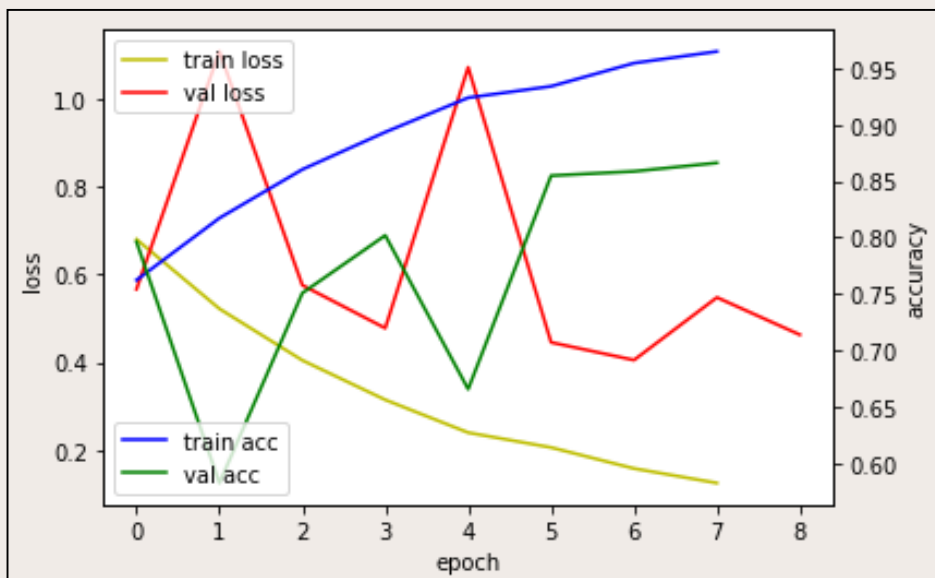
컬럼별 모델 선정 결과

	Tokenizer	lr_rate	F1-score
혐오	KcELECTRA	0.001	0.685
성별	KcBERT	0.001	0.660
이념/지역	KcELECTRA	0.0005	0.346
인간존엄성	KcBERT	0.001	0.345

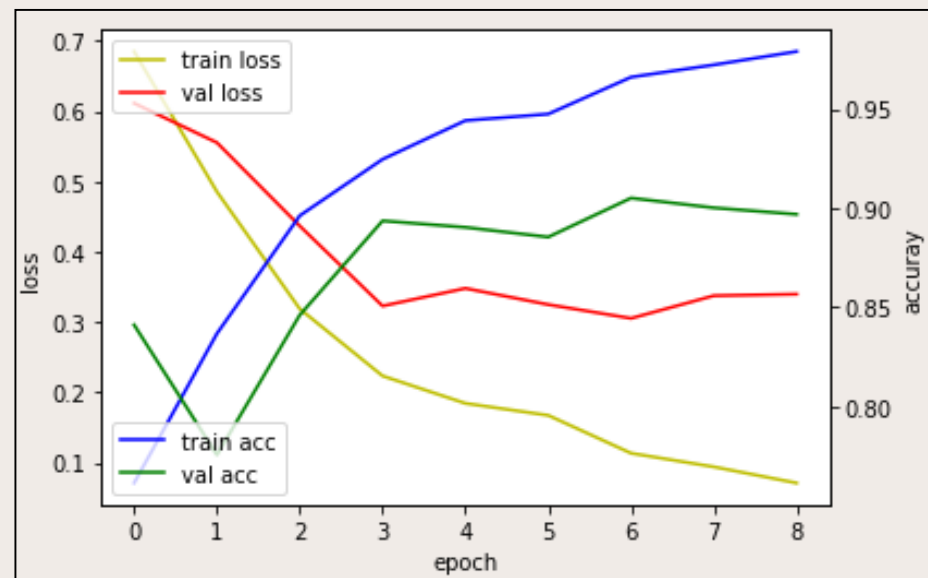


컬럼별 Train결과

혐오



성별



F1-score	0.6846
accuracy	0.8430
recall	0.7238
precision	0.6495

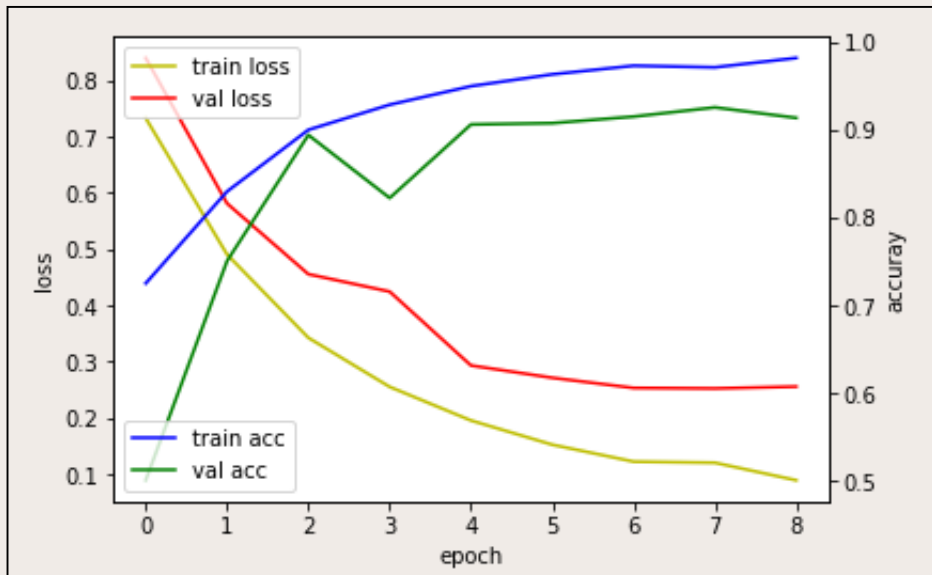
Evaluate
Result

F1-score	0.6595
accuracy	0.8923
recall	0.7750
precision	0.5740

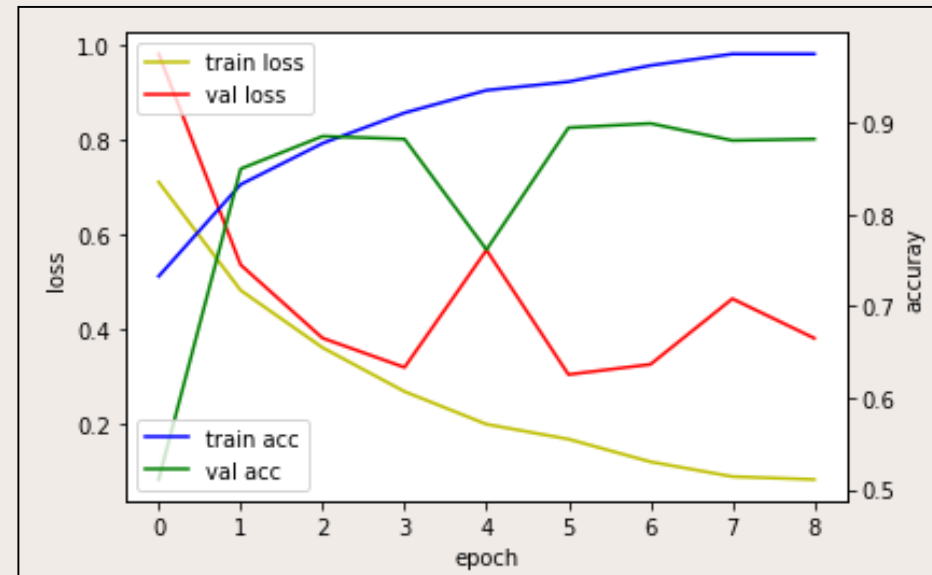


컬럼별 Train결과

이념/지역



인간존엄성



F1-score	0.3458
accuracy	0.9024
recall	0.5227
precision	0.2584

Evaluate
Result

F1-score	0.3454
accuracy	0.8385
recall	0.4418
precision	0.2853



각 댓글별 혐오 확률

test_data: 택시운전사

```
test_data[['text', 'Pr']].sample(10)
```

	text	Pr
12361	반전없는 범작이다. 독일기자의 눈에 비친 모습이 저게 다는 아닐 것이다.	0.032005
19763	너무 슬펐어요... 전두환은 왜 아직도 살아있을까	0.041819
47493	광주는 그.... 그 하나의	0.870661
31110	재밌게 잘 봤어요. 실화라고 하는데 안타깝더라고요...	0.050618
47729	광주 고만좀 울거먹어라 . 지겹다	0.556835
6930	와 별10개론 부족...	0.004455
33874	아라아아라아두웅어언어 ㅈ	0.377984
29575	넘 실속있고 좋았어요아픈역사 공부도 많이 했네요연기력 정말 짱!	0.165678
27847	어설픈 연출과 억지 감동을 짜내는 부분이 많아 몰입도가 떨어졌다. 하지만 작품의 취...	0.000883
45351	꼭 보아야 할 영화입니다	0.035059

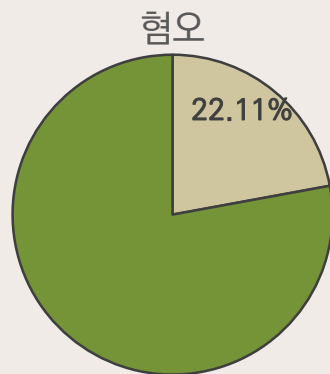
Test set 각 댓글들의 혐오성을 판단

0~1 사이의 값으로, 1에 가까울수록
→ 혐오성이 짙은 댓글



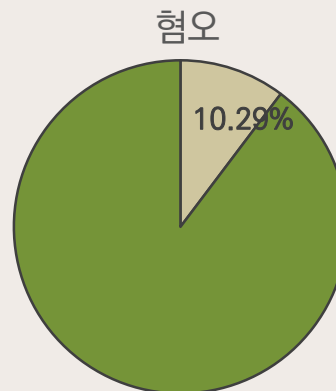
전체 댓글의 혐오/비혐오 비율

캡틴마블



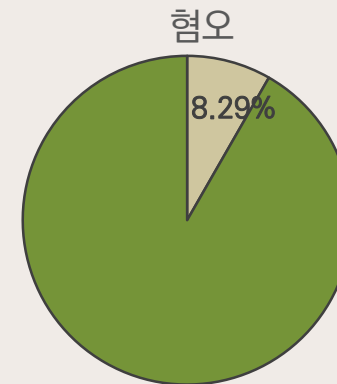
■ 혐오 ■ 비혐오

택시운전사



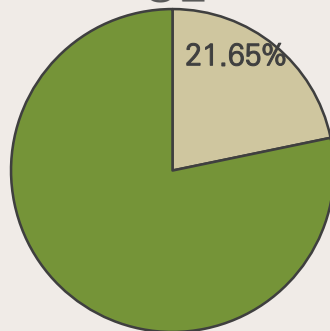
■ 혐오 ■ 비혐오

공작



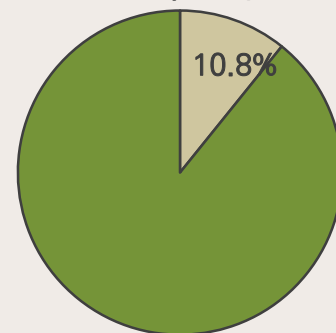
■ 혐오 ■ 비혐오

성별



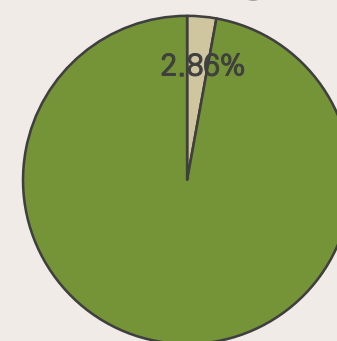
■ 혐오 ■ 비혐오

이념/지역



■ 혐오 ■ 비혐오

인간존엄성



■ 혐오 ■ 비혐오



실평점 예측

```
test_data['혐오/비혐오'] = pred_test2
test_data

#혐오가 아닌 댓글만
test_data_real = test_data[test_data['혐오/비혐오'] == 0]

## 실평점 예측 ##
#실평점예측 = 전체평점/댓글수
real_score= test_data_real.score.sum()/len(test_data_real)
print("이 영화의 네이버 평균 평점은 % 0.2f 점입니다." % test_data['score'].mean())
print("이 영화의 실평점은 % 0.2f 점입니다." %(real_score))
```

캡틴마블

네티즌 평점



6.75

실 평점



6.83

택시운전사

네티즌 평점



9.05

실 평점



9.07

공조

네티즌 평점



8.27

실 평점



7.21



댓글 혐오 판별기

혐오 댓글 예시

군대나 가고 이런 말 해라 피해망상 환자들아

군대나 가고 이런 말 해라 피해망상 환자들아
1/1 [=====] - 1s 1s/step
위 댓글이 혐오일 확률은 87.54 % 입니다.

비혐오 댓글 예시

영화 보다가 너무 졸려서 그냥 나왔음

영화 보다가 너무 졸려서 그냥 나왔음
1/1 [=====] - 2s 2s/step
위 댓글이 혐오일 확률은 0.57 % 입니다.

Concl.





국가인권위원회

독자가 많은 매체를 통한
혐오 표현일수록 커지는
해악성

불법 온라인 혐오표현에 대응하기 위한 행동준칙

IT기업들이 혐오표현을 삭제,
차단하고 신고



검색 매체로서 네이버의 독보적인 점유율
혐오 표현을 선제적으로 통제하는 수단을 도입할 필요성
‘혐오 댓글 판별기’를 통한 선제적 차단 가능



제언

NAVER 영화

영화홈

상영작·예정작

현재 상영영화

개봉 예정영화

예고편

영화랭킹

예매

평점·리뷰

다운로드

인더극장

보기음선

기본보기

넓게보기

관람객 8.26

기자·평론가 6.00

네티즌

개요

감독

출연

등급

흥행

성별·나이별 관람

예매하기

주요정보

네티즌 평점

인질

별점을 선택하세요

어떤 점이 좋았나요? 감상포인트를 추천해주세요! 중복 선택 가능

연출

연기

스토리

영상미

OST

액션 지수를 선택해 주세요

높음

보통

낮음

광주는 그...하나의...나

16 / 1000

감상평에 스포일러가 포함되어있나요?

있음

없음

취소

확인

네이버 영화



혐오 표현이 감지되어 댓글을 등록할 수 없습니다.
사유: 이념/지역 혐오 맥락 탐지 (89%)

확인



제언



절대적인 데이터 부족

- 해당 영화평의 약 21%만 train set으로 사용됨 (4,476개)
- 선행연구에서는 20만 개 이상의 충분한 데이터를 사용



‘평점 알바’가 높이는 평점 문제

- 본 알고리즘만으로는 의도적으로 높은 평점을 주는 댓글을 필터링하기 어려움
- 높은 평점을 위해 고의적으로 높은 점수를 주는 리뷰를 필터링하는 보완 모델 필요



데이터 불균형

- 혐오:비혐오 비율 3:1
- 성별 637개, 인간존엄성 427개 vs 이념/지역 271개

Column	f1 score
혐오	0.685
성별	0.346
이념/지역	0.660
인간존엄성	0.345



“딥러닝 기술을 활용한 차별 및 혐오 표현 탐지: 어텐션 기반 다중 채널 CNN모델링”, 이원석 외 1인, 한국정보통신학회지 Vol 24, No.12: 1595~1603, 2020

“혐오표현 리포트”, 국가인권위원회, 2019

<https://chioni.github.io/posts/electra/>

<https://github.com/whjzsy/AMCNN>

https://support.google.com/youtube/answer/2801939?hl=ko&ref_topic=9282436



CNN모델링을 활용한 혐오 댓글 판별기

감사합니다.