악성 댓글 탐지 모델

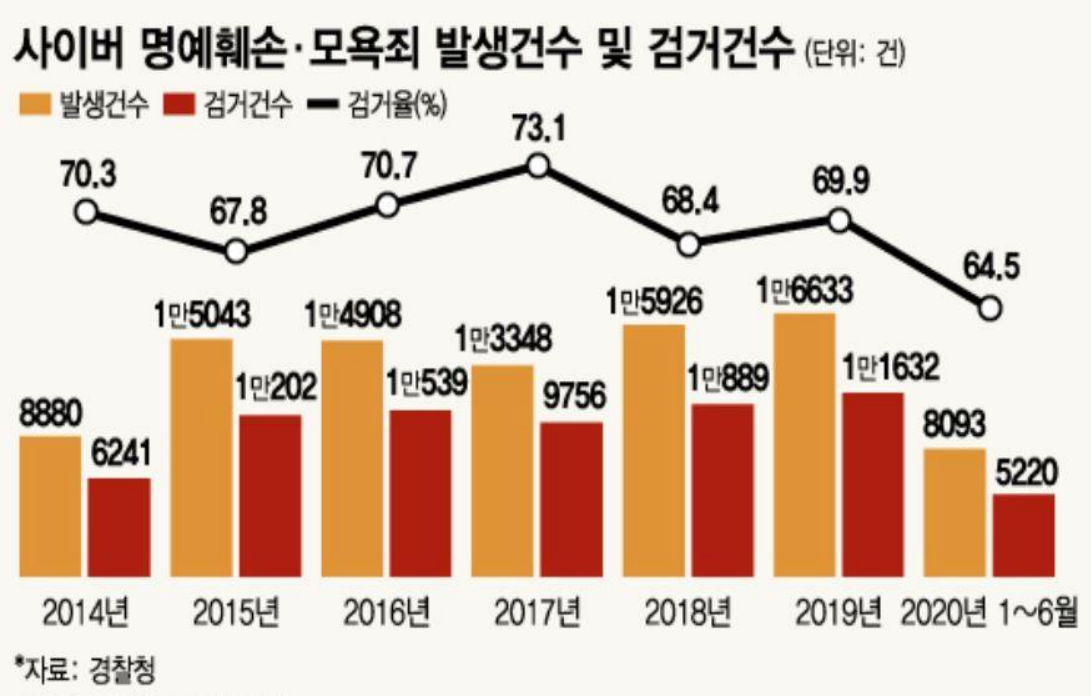
AI 10 우경화

Contents

01.

프로젝트 개요

1-1. 프로젝트 배경



- ✓ 폭언 및 욕설, 악의적인 비방, 허위사실 유포
- ✓ 개인 SNS, 유튜브와 같은 콘텐츠 댓글 테러
- ✓ 매해 늘어나고 있는 악성 댓글 사건들

1-2. 프로젝트 목적

폭언 및 욕설, 악의적인 비방, 혐오 발언 등 **악성 댓글을 탐지**

올바른 댓글 매너 및 자유로운 소통 문화 만들기

02.

데이터 소개 및 가설

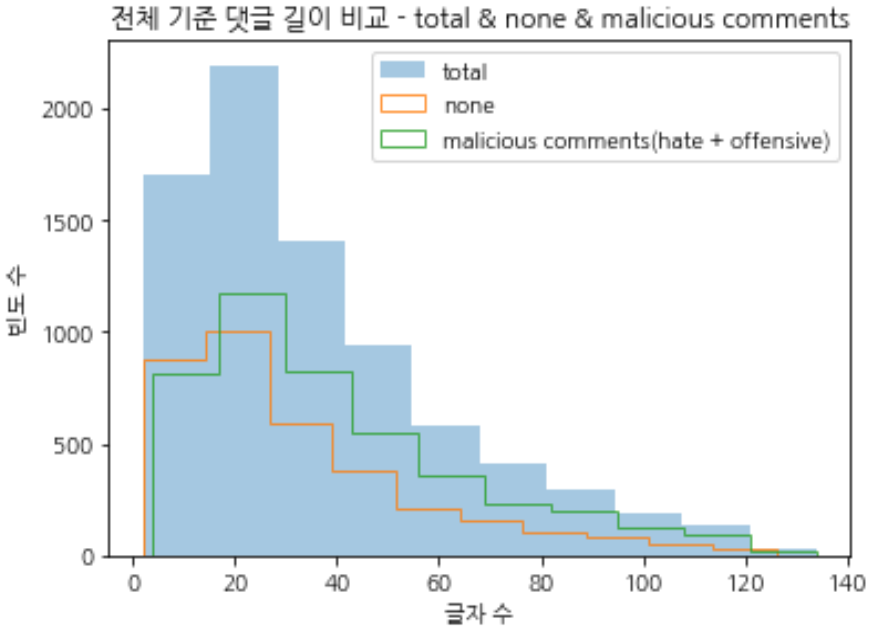
2-1. 데이터 소개

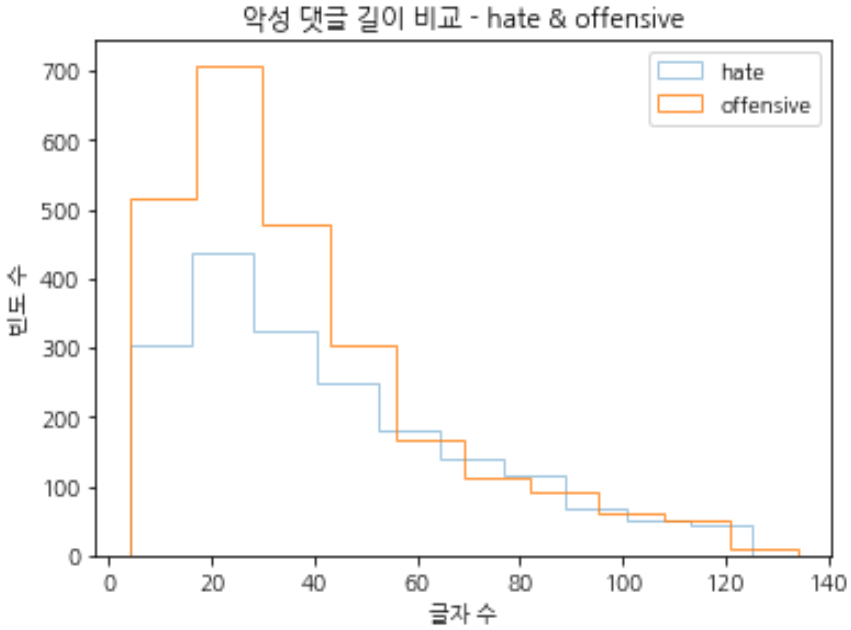
출처 : https://github.com/kkobooc/NLP_KoreanHateSpeech

	comments	contain_gender_bias	bias	hate	news_title
0	(현재 호텔주인 심정) 아18 난 마른하늘에 날벼락맞고 호텔망하게생겼는데 누군 계속...	False	others	hate	"밤새 조문 행렬...故 전미선, 동료들이 그리워하는 따뜻한 배우 [종합]"
1	....한국적인 미인의 대표적인 분...너무나 곱고아름다운모습...그모습뒤의 슬픔을...	False	none	none	"'연중' 故 전미선, 생전 마지막 미공개 인터뷰...환하게 웃는 모습 '먹먹'[종합]"
2	...못된 넘들...남의 고통을 즐겼던 넘들..이젠 마땅한 처벌을 받아야지..,그래...	False	none	hate	"[단독] 잔나비, 라디오 출연 취소→'한밤' 방송 연기..비판 여론 ing(종합)"
3	1,2화 어설펐는데 3,4화 지나서부터는 갈수록 너무 재밌던데	False	none	none	"'아스달 연대기' 장동건-김옥빈, 들끓는 '욕망커플'→눈물범벅 '칼끝 대립'"
4	1. 사람 얼굴 손톱으로 긁은것은 인격살해이고2. 동영상이 몰카냐? 메걸리안들 생각...	True	gender	hate	[DA:이슈] '구하라 비보' 최종범 항소심에 영향?...법조계 "'공소권 없음' 아냐"

- comments : 댓글
- hate
  - none : 욕설이나 모욕이 내포되지 않은 일반 댓글(<-> 악성댓글)
  - hate : 욕설 또는 강한 혐오표현, 비난이 들어간 악성 댓글
  - offensive : hate만큼 모욕적이진 않지만 공격적이고 사람을 불쾌하게 만드는 댓글

2-2. 가설 설정


전체 기준 댓글 길이 비교 - total & none & malicious comments
total
none
malicious comments(hate + offensive)
빈도 수
글자 수
0
500
1000
1500
2000
0
20
40
60
80
100
120
140


악성 댓글 길이 비교 - hate & offensive
hate
offensive
빈도 수
글자 수
0
100
200
300
400
500
600
700
0
20
40
60
80
100
120
140

03.

모델 프로세싱

3-1. 데이터 전처리

- ✓ 결측치 확인 후 제거
- ✓ 기본 문자표 제거
- ✓ 한글, 영어, 숫자를 제외한 모든 문자열 제거
- ✓ 중복되는 문자 제거
- ✓ 프로젝트에 필요한 컬럼('comments', 'hate')만 추출

3-2. 모델링

	comments	hate
0	현재 호텔주인 심정 아18 난 마른하늘에 날벼락맞고 호텔망하게생겼는데 누군 계속 추모받네	1
1	한국적인 미인의 대표적인 분너무나 곱고아름다운모습그모습뒤의 슬픔을 미처 알지못했네요ㅠ	0
2	못된 넘들남의 고통을 즐겼던 넘들이젠 마땅한 처벌을 받아야지그래야 공정한 사회지심은...	1
3	12화 어설펐는데 34화 지나서부터는 갈수록 너무 재밌던데	0
4	1 사람 얼굴 손톱으로 긁은것은 인격살해이고2 동영상이 몰카냐 메걸리안들 생각이 없노	1

- hate 컬럼을 0과 1로 이진 분류
  - 0 : none (일반 댓글)
  - 1 : offensive , hate (악성 댓글)

3-2. 모델링

BERT

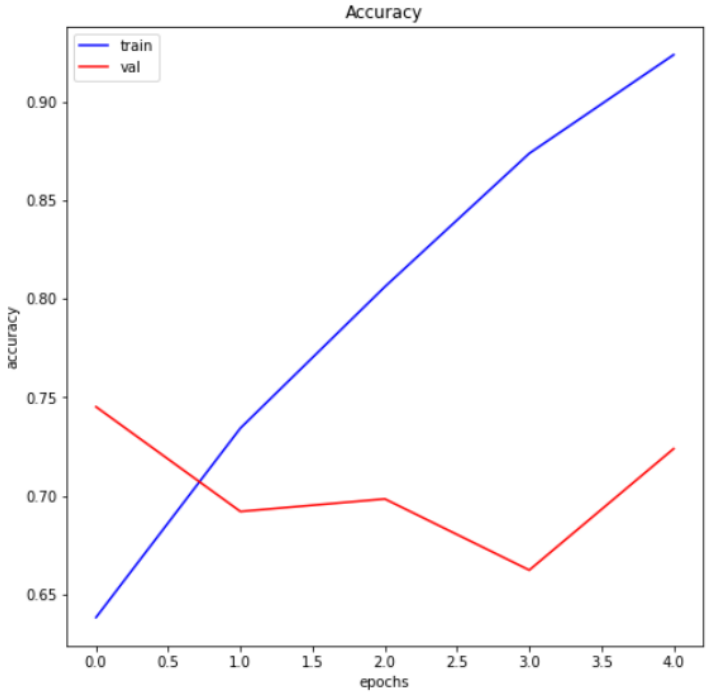
(Bidirectional Encoder Representations from Transformers)

✓ Google에서 개발한 자연어 처리 기술

✓ Transformer 구조를 응용한 모델

✓ 대문자/소문자에 따라 다른 단어로 인식함

3-3. 모델 학습 결과



val accuracy : 0.7239915074309978
val f1_score : 0.7750865051903113

3-3. 모델 학습 결과

	Comment	multilingual_model
0	ㅋㅋ 그래도 조아해주는 팬들 많아서 좋겠다 ㅠㅠ 니들은 온유가 안만져줌 ㅠㅠ	1
1	둘다 넘 좋다행복하세요	0
2	근데 만원이하는 현금결제만 하라고 써놓은집 우리나라에 엄청 많은데	0
3	원곡생각하나도 안나고 러블리즈 신곡나온줄 너무 예쁘게 잘봤어요	0
4	장현승 얘도 참 이젠 짠하다	1
...	...	...
969	대박 게스트 꼭 봐야징 컨셉이 바뀌니깐 재미지넹	0
970	성형으로 다 뜯어고쳐놓고 예쁜척 성형 전 니 얼굴 다 알고있다 순자처럼 된장냄새 나게 생겼더만	1
971	분위기는 비슷하다만 전혀다른 전개던데 무슨ㅋㅋㄱ 우리나라사람들은 분위기만 비슷하면 다 표절이래 그럼 클래식계열도 다 표절이고 재즈계열도 다 표절이게	0
972	입에 손가락이 10개 있으니 징그럽다	1
973	난 조보아 이뻐서 보는데 백종원 관심무	0

04.

결론 및 향후 개선사항

4-1. 결론 및 향후 개선사항

✓ 전반적으로 일반댓글은 0, 악성댓글은 1로 잘 분류됨

✓ 인터넷 댓글이기 때문에, 오타가 많으며 맞춤법이 정확하지 않아 토큰화가 제대로 되지 않는 경우를 개선

✓ 인터넷 유행어나 신조어 등이 제대로 분석되지 않는 부분 개선

감사합니다