머신러닝을 활용한 한국어 악성 댓글 종류 판별 시스템

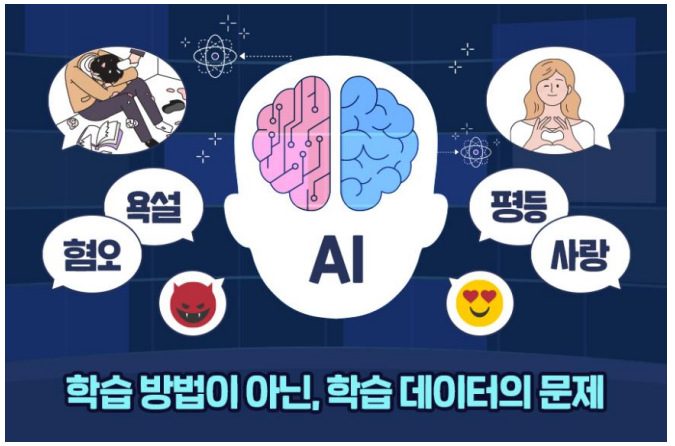
기계학습 입문 16조: 정상헌(2019313145), 손정용(2018312407), 정은채(2023312807)

목차

1. 주제 선정 배경

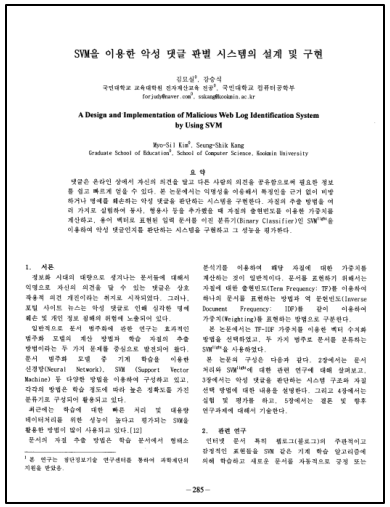
댓글 / 악성 댓글

악성 댓글로 인한 사회, 경제적 비용 발생



댓글이 학습 데이터로 이용될 때 윤리 문제 발생

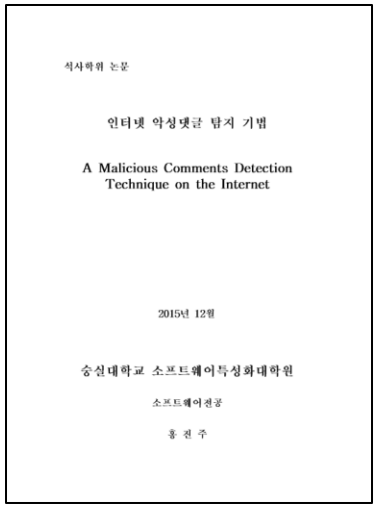
2. 선행연구 소개



SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현

A Design and Implementation of Malicious Web Log Identification System by Using SVM

[1]



석사학위 논문

인터넷 악성댓글 탐지 기법

A Malicious Comments Detection Technique on the Internet

2015년 12월

숭실대학교 소프트웨어특성화대학원

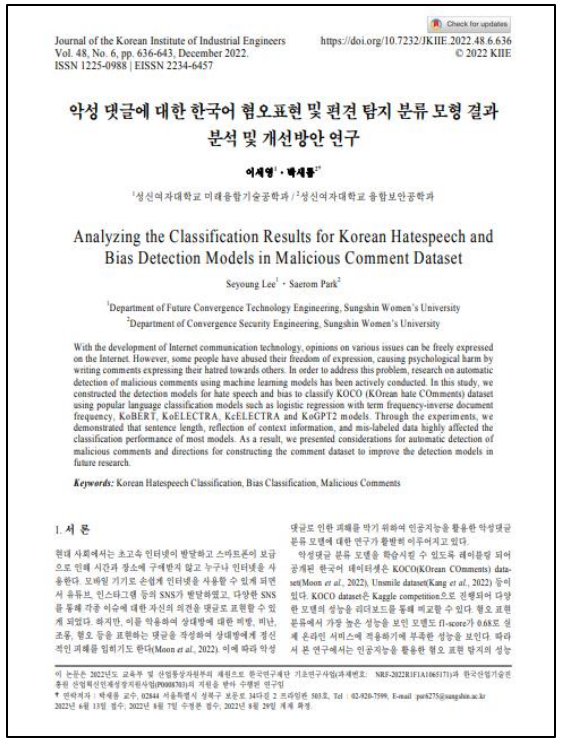
소프트웨어전공

[2]

기계 학습을 이용한 악성 댓글 판별 시스템

Discrimination System for Abusive Comments using Machine Learning

[3]



Journal of the Korean Institute of Industrial Engineers
Vol. 48, No. 6, pp. 636-643, December 2022.
ISSN 1225-0988 | EISSN 2234-6457

https://doi.org/10.7232/JKIIE.2022.48.6.636
© 2022 KIIE

악성 댓글에 대한 한국어 혐오표현 및 편견 탐지 분류 모형 결과 분석 및 개선방안 연구

Analyzing the Classification Results for Korean Hatespeech and Bias Detection Models in Malicious Comment Dataset

[4]

-> 머신러닝을 활용한 악성 댓글 분류 선행 연구

악성 댓글 유무 판별, 이진 분류

SVM 알고리즘

-> LR, 딥러닝을 활용한 선행 연구

혐오 정도, 편견, 차별에 따른 다중 분류

Logistic Regression, 딥러닝

2. 선행연구 소개

선행연구 [1], [2], [3] 과 달리 악성 댓글의 유형을
[출신차별/외모차별/정치성향차별/혐오욕설/연령차별/성차별/인종 차별/종교차별]
로 다중 분류하여 예측 프로젝트 진행

1. 선행연구 [4]에서 딥러닝과 비교했을 때에도 좋은 성능을 보인 **Logistic Regression**

2. 확률적인 접근을 기반으로 하여 속도가 빠르다는 장점이 있는 **Naïve Bayes**

3. 선행연구에서 잘 사용하지 않은 트리 기반 앙상블 학습 알고리즘 **XGBoost**

사용한 데이터 셋

- K-MHaS (한국 온라인 뉴스 댓글)

Train: 78,977
Valid: 8,776
Test: 21,939

=> Total: 109,692

- Multi-Label 데이터 셋
- 수집기간: 2018.01 ~ 2020.06

악성 댓글 종류

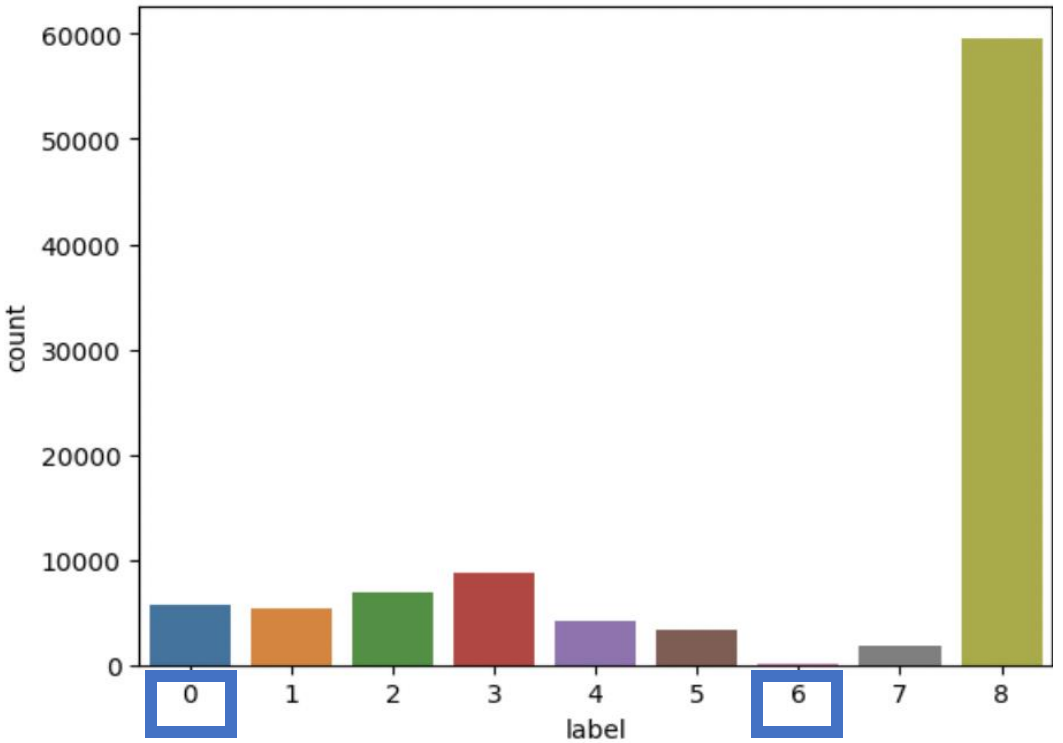
- 출신차별 (0)
- 외모차별 (1)
- 정치성향 차별 (2)
- 혐오욕설 (3)
- 연령차별(4)
- 성차별 (5)
- 인종차별 (6)
- 종교차별 (7)
- 해당사항 없음 (8)

=> 총 9개 클래스

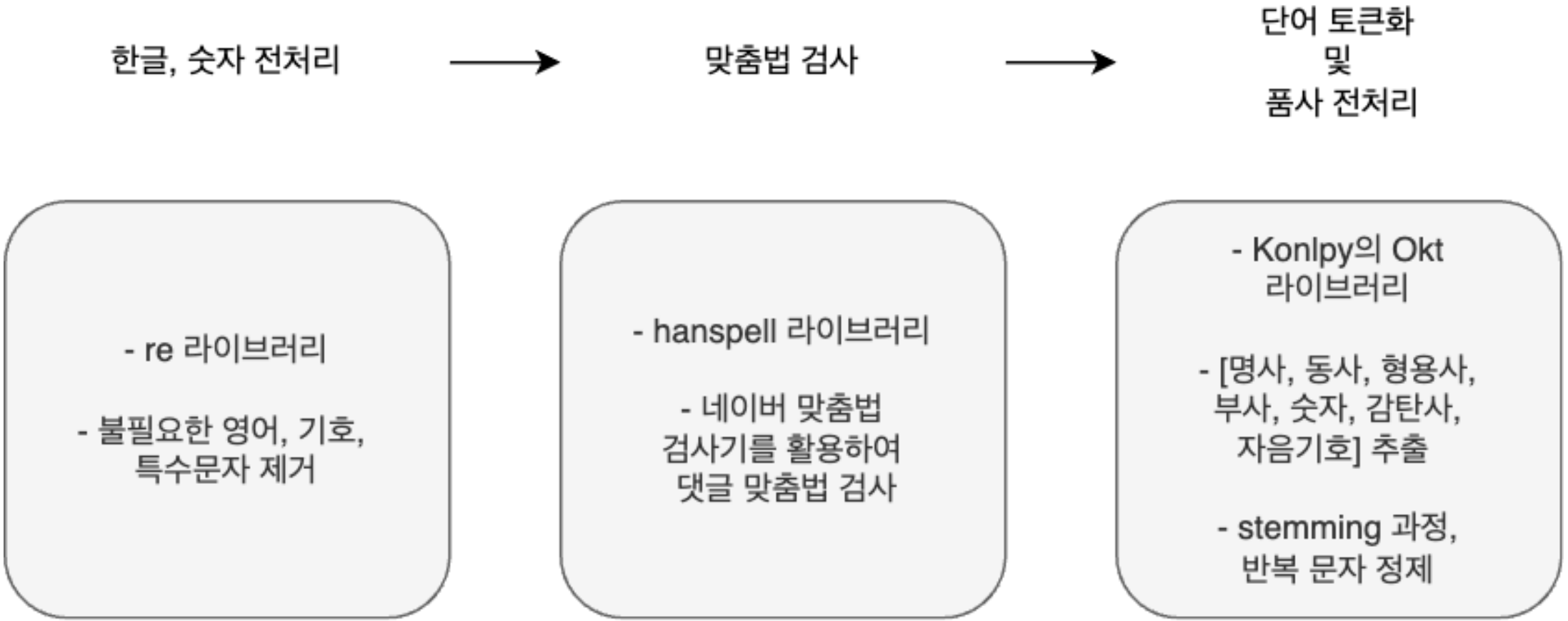
레이블 전처리

레이블 개수가 2개 이상인 데이터 제거 & 인종차별(6), 출신차별(0) 클래스 결합

	document	label
0	자한당틀딱들.. 악플질 고만해라.	2,4
1	정치적으로 편향된 평론한은 분은 별로...	8
2	적당히좀 쳐먹지.그랬냐??? 안그래도 문재인 때문에 나라 엉망진창인데...	2
3	안서는 아재들 풀발기 ㅋㄲㅋ	4
4	우와 ㅋ 능력자	8
5	맛녀석 콩트보다 약했음맛녀석 애청자로써 70%실력발휘	8
6	주영훈 솔직히 호감임 잉꼬부부로 소문났잖아	8
7	이게주간아이돌이랑머가달라...	8
8	아오 슈박 회사생활도 졑깥고 돈벌기 힘들어 죽겠구만 뭔 저딴것들 자꾸 tv나와서 사...	3
9	문재인 하는게 뭐 별거있냐?ㅂㅅㅅㅋ가 하는짓인데 어련하겠어.ㅋㅋㅋ	2,3



텍스트 전처리


한글, 숫자 전처리
맞춤법 검사
단어 토큰화
및
품사 전처리
- re 라이브러리
- 불필요한 영어, 기호,
특수문자 제거
- hanspell 라이브러리
- 네이버 맞춤법
검사기를 활용하여
댓글 맞춤법 검사
- Konlpy의 Okt
라이브러리
- [명사, 동사, 형용사,
부사, 숫자, 감탄사,
자음기호] 추출
- stemming 과정,
반복 문자 정제

텍스트 전처리

	document	label	spelled	tokens
0	정치적으로 편향된 평론한은 분은 별로...	7	정치적으로 편향된 평론 한은 분은 별로	정치 편향 되다 평론 은 분 별로
1	적당히좀 쳐먹지.그랬냐??? 안그래도 문재인 때문에 나라 엉망진창인데...	2	적당히 좀 처먹지 그랬냐 안 그래도 문재인 때문에 나라 엉망진창인데	적당하다 좀 처 먹다 그렇다 안 그래도 문재인 때문 나라 엉망 진창
2	안서는 아재들 풀발기 ㅋㄲㅋ	4	안 서는 아재들 풀 발기 ㅋㄲㅋ	안 서다 아 재 풀 발기 ㅋㄲㅋ
3	우와 ㅋ 능력자	7	우와 ㅋ 능력자	우와 ㅋ 능력자
4	맛녀석 콩트보다 약했음맛녀석 애청자로써 70%실력발휘	7	맛 녀석 콩트보다 약했음만 녀석 애청자로써 70실력 발휘	맛 녀석 콩트 약하다 음 녀석 애 청자 로써 70 실력 발휘
...	...	...	...	...
95956	신천지들 소원대로 전부 하느님 곁으로 보내주세요	6	신천지들 소원대로 전부 하느님 곁으로 보내주세요	신천지 소원 전부 하느님 곁 보내다
95957	댓글보니 다 언니사랑해요 언니언니 거리네확실히 뷰티유튜버라 여자들 댓글이 대부분이네	5	댓글 보니 다 언니 사랑해요 언니 언니 거리네 확실히 뷰티 유튜버라 여자들 댓글이 ...	댓글 보다 다 언니 사랑 하다 언니 언니 거리 확실하다 뷰티 유튜버 여자 댓글 대부분
95958	중국이란 나라의 수준을 고스란히 보여주네...	7	중국이란 나라의 수준을 고스란히 보여주네	중국 나라 수준 고스 란 히 보여주다
95959	김새롬같은 철부지를 아내로 받아드린게 잘못ㅠㅠ 힘내세요ㅠㅠ	7	김새롬 같은 철부지를 아내로 받아들인 게 잘못ㅠㅠ 힘내세요ㅠㅠ	김새롬 같다 철부지 아내 받아들이다 게 잘못 ㅠㅠ 힘내다 ㅠㅠ
95960	송가인은 미운 우리새끼에 나와야하는거아니냐?	7	송가인은 미운 우리 새끼에 나와야 하는 거 아니냐	송가 미우다 우리 새끼 나오다 하다 거 아니다

텍스트 수치화 방법

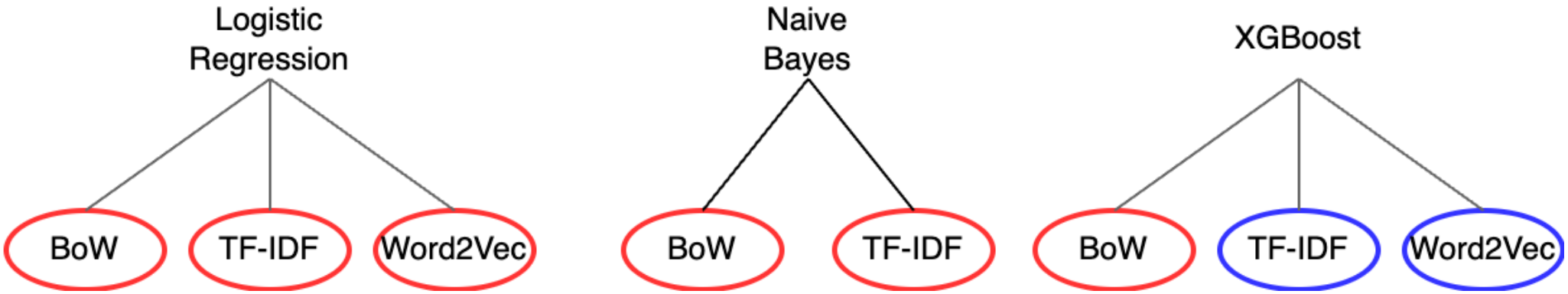
BoW | 단어들의 순서는 전혀 고려하지 않고, 단어들의 출현 빈도(frequency)에 만 집중하는 텍스트 데이터의 수치화 표현 방법

TF-IDF | 단어의 빈도와 역 문서 빈도를 사용하여 각 단어 들마다 중요한 정도를 가중치로 주는 방법

Word2Vec | 단어 간의 유사성을 기반으로 단어를 벡터로 표 현하는 방법 두 가지 주요 방법으로 "CBOW (Continuous Bag of Words)"와 "Skip-gram”으로 구분한다.

‘CBOW’: 주변 단어들을 사용해서 특정 단어를 예측
‘Skip-gram’: 대상 단어로 주변 단어를 예측

4. 모델 방법론



Random Search

Random search, Grid search 시간소요

→ BoW 방식 파라미터 범위 고려해 결정

-> 총 8개의 모델

평가지표: Accuracy, Precision, Recall, F-1 Score

다중 분류 문제: 클래스 별 Confusion Matrix 값 -> Macro 평균

오버 샘플링

데이터의 레이블 불균형 문제 해결을 위해 BoW 임베딩 방식에서
오버 샘플링 적용 데이터 vs 기존 데이터 간단한 성능 비교

오버 샘플링 기법을 활용한 데이터보다 기존의 데이터를 사용할 때 모델의 성능이 더 좋게 나옴

=> 본 프로젝트에서는 **기존의 데이터 셋**을 그대로 사용하여 진행

모델 결과 분석

BoW Embedding			
	Logistic Regression	Naive Bayes	XGBoost
Acc	0.776	0.750	0.790
Precision	0.743	0.652	0.757
Recall	0.579	0.603	0.622
F1 Score	0.642	0.618	0.675

TF-IDF Embedding			
	Logistic Regression	Naive Bayes	XGBoost
Acc	0.774	0.731	0.781
Precision	0.746	0.730	0.741
Recall	0.578	0.438	0.609
F1 Score	0.643	0.513	0.660

Word2Vec Embedding		
	Logistic Regression	XGBoost
Acc	0.633	0.654
Precision	0.570	0.553
Recall	0.168	0.249
F1 Score	0.178	0.299

- 기본적으로 XGBoost 모델에서 가장 좋은 성능을 보임
- Word2Vec 임베딩 기법에서는 모두 좋지 않은 성능을 나타냄

5. 결과 분석

최적의 모델

-> BoW 임베딩에서 XGBoost 알고리즘을 이용해 학습한 모델 (Acc: 0.790)

하이퍼 파라미터 조합

-> {'tree_method': 'hist', 'n_estimators': 800, 'max_depth': 10, 'learning_rate': 0.1}

BoW Embedding			
	Logistic Regression	Naive Bayes	XGBoost
Acc	0.776	0.750	0.790
Precision	0.743	0.652	0.757
Recall	0.579	0.603	0.622
F1 Score	0.642	0.618	0.675

실제 예시 댓글 모델 성능

[댓글1]

'짱개는 어떻게 저렇게 하루종일 민폐만 끼치지 진짜ㅋㅋ'

```
출신차별 & 인종차별(0): 99.32%
외모차별(1): 0.05%
정치성향차별(2): 0.03%
혐오욕설(3): 0.13%
연령차별(4): 0.08%
성차별(5): 0.02%
종교차별(6): 0.01%
해당사항없음(7): 0.35%
* 최종 예측 유형: 출신차별 & 인종차별(0) 99.32%
```

[댓글2]

'요즘 말도 많고 탈도 많고 ㅈㄹ들인지ㅉㅉ'

```
출신차별 & 인종차별(0): 0.12%
외모차별(1): 0.03%
정치성향차별(2): 0.1%
혐오욕설(3): 97.85%
연령차별(4): 0.33%
성차별(5): 0.11%
종교차별(6): 0.02%
해당사항없음(7): 1.44%
* 최종 예측 유형: 혐오욕설(3) 97.85%
```

[댓글3]

'부럽다 진짜ㅠㅠ'

```
출신차별 & 인종차별(0): 0.68%
외모차별(1): 0.52%
정치성향차별(2): 0.63%
혐오욕설(3): 0.76%
연령차별(4): 0.23%
성차별(5): 0.35%
종교차별(6): 0.05%
해당사항없음(7): 96.78%
* 최종 예측 유형: 해당사항없음(7) 96.78%
```

데이터 분석 의의

- 선행연구와 달리, 이 프로젝트에서는 다중 분류를 하여 악성 댓글의 여부 뿐만 아니라 어떤 유형의 악성 댓글인지까지도 판별할 수 있는 모델 구현

- 관련 선행 연구에서 잘 사용하지 않았던 트리 기반 알고리즘인 XGBoost를 사용하여 최고 성능 달성

- 실제 데이터 셋을 활용하여 79% 정도의 유의미한 성능 도출

활용 방안

1. 악성 댓글 감소

- 작성한 댓글의 종류와 횟수를 해당 아이디에 할당하여 아이디별로 표식시스템을 만드는 것
- 댓글을 입력한 후 작성을 누리기 전에 미리 경고 알림을 하는 시스템을 만드는 것

2. 학습 데이터 구축에 활용

- 데이터 학습 전, 비윤리성을 띠는 요소를 제거 or 옳지 않은 표현들을 하지 않도록 학습

한계점

- Colab 환경에서 프로젝트를 수행하여 자원적 제약이 있었음

- 동일한 실험을 하기 위한 제한 조건을 통일하지 못함

(XGBoost에서 BoW와 달리, TF-IDF 와 Word2Vec에서 랜덤서치 불가능)

-> 다양한 모델과 넓은 범위에서 하이퍼 파라미터 탐색을 진행하여 최종 모델의 성능을 향상 기대

- 사용한 데이터 셋에서 클린한 댓글의 수가 압도적으로 많음
- 전처리 과정에서 걸러내지 못한 단어, 표현 존재 가능성 O

-> 더욱 정교한 전처리 과정을 통해 모델 성능 향상 기대