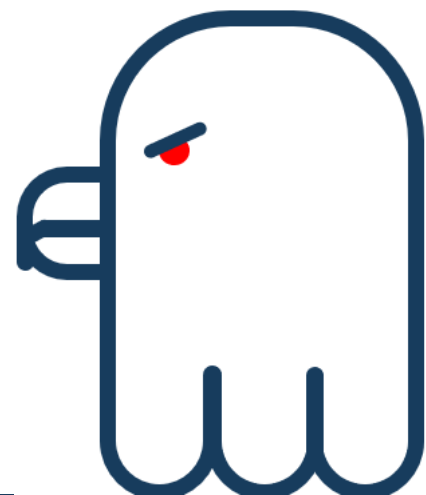


매의 눈 - 악성 댓글 분류 시스템



7th BOAZ BIGDATA CONFERENCE

김 수 연

김 유 경

김 현 중

박 정 우

최 하 은

INDEX



주제 선정 배경



데이터 설명



모델 구현



결론

I

주제 선정 배경

주제 선정 배경



[제천 부시장 “유족 소방관들 악성 댓글에 큰 상처”](#)

뉴스1 | 2017.12.27. | 네이버뉴스 | [🔗](#)

© News1 자제 호소...입에 담지 못할 내용 많아 유연비어도 난무



[수능 최고령 응시자에 '악장' 댓글 단 사람들](#)

헤럴드경제 | 2017.11.23. | 네이버뉴스 | [🔗](#)



[\[초점IS\] 종현, 도 넘은 '무개념' 악플..팬들이 나선다](#)

일간스포츠 | 2017.12.20. | 네이버뉴스 | [🔗](#)

종현은 생전 악플로 힘들어했다. 상당수의 연예인들이 악플에 시달 소속사에서 꾸준히 악플러에게 경고했지만, 악플의 뿌리를 완전히 통 속에서 눈을 감은 뒤에도...

“알몸으로 사망했네. 부끄럽다”

“시험장에서 틀니소리 조심해주세요.”

“성형수술 후유증인 것 같음”

“사람을 구하지도 못하면서 쇼하러 출동했냐.”

“늙어서 주책은..ㅍ..ㅍ”

“군대갈 때가 되니 두려운거지”

“애도하는 마음이 싹 달아난다. 유가족 갑질 장난 아니네.” “하나 제끼고 갑니다^^”

“이제 김연탄이노?”

주제 선정 배경



[제천 부시장 “유족 소방관들 악성 댓글에 큰 상처”](#)

뉴스1 | 2017.12.27. | 네이버뉴스 |

© News1 자제 호소...입에 담지 못할 내용 많아 유연비어도 난무



[수능 최고령 응시자에 '악장' 댓글 단 사람들](#)

헤럴드경제 | 2017.11.23. | 네이버뉴스 |



[\[초점IS\] 종현, 도 넘은 '무개념' 악플..팬들이 나선다](#)

일간스포츠 | 2017.12.20. | 네이버뉴스 |

종현은 생전 악플로 힘들어했다. 상당수의 연예인들이 악플에 시달 소속사에서 꾸준히 악플러에게 경고했지만, 악플의 뿌리를 완전히 통 속에서 눈을 감은 뒤에도...

익명성을 무기로 한

도넘은 악플에 대한 제재가 필요함



사용자와 함께 만드는 댓글 문화

1

내가 보고 싶지 않은
댓글이 있다면 바로 **접기요청**하여
해당 댓글을 접을 수 있어요!

2

접힌 댓글을 보고 싶을 경우 **펼쳐서**
그 내용을 확인할 수 있어요!

3

다수의 요청으로 자동접힌 댓글에 대해
추가 평가도 가능합니다 :)

1

mega****

너희들이 하는 짓이 뭐 그렇지...꼴도 보기 싫은 녀석들
한대 맞아야 정신'

2017-05-25 16:3

접기 요청

답글 32

1221

30

↓

2

mega****

ⓘ 사용자 요청으로 접힌 댓글입니다

2017-05-23 11:02 | [내용 보기](#)

↓

3

mega****

너희들이 하는 짓이 뭐 그렇지...꼴도 보기 싫은 녀석들
한대 맞아야 정신차리지.

2017-05-23 11:02 | [닫기](#)

답글 32

1221

30

다수가 접기 요청한 댓글입니다.
[위 댓글에 대한 의견은?](#)

접기유지

펼침 요청

제한적인 자동 필터링

심한 욕설의 경우는 작성 불가능

이용자들의 신고

① 다수의 사용자의 요청에 따라 자동으로 접힌 댓글입니다.



악성 댓글 차단 기술 현황

× 댓글

✓

Default Keywords

부적절한 댓글 숨기기

Hide comments that contain words or phrases often reported as offensive.

Custom Keywords

Add keywords, separated by commas

Comments that contain any of the words or phrases above will be hidden.

‘부적절한 댓글 숨기기 (Hide Offensive Comments)’

게시물과 라이브 방송에서 불쾌한 댓글을 막는 필터를 도입

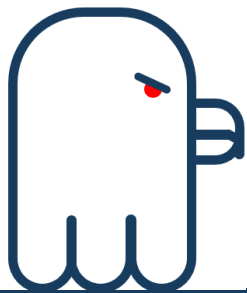
AI 적용

Deep Text 텍스트 분류 기술 기반의 댓글 필터링

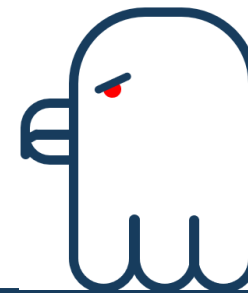
사용자 지정단어 입력하면 자동 필터링



“ 현재의 자동 필터링 시스템이 **한계**를 갖고 있는 상황.
이를 보완할 수 있는 필터링 시스템 시도 ”



Project



매의 눈 _

악성 댓글 분류 시스템


||

데이터 설명

데이터 수집


NAVER 뉴스 | TV연예 | 스포츠 | 뉴스스탠드 | 날씨

뉴스홈 속보 정치 경제 사회 생활/문화 세계 IT/과학 오피니언 포토 TV **랭킹뉴스**


01.13 (토)  서울 -3°C 주요뉴스 ▶ CPU 보안패치마저 결함...갑자기 PC재부팅 현상


랭킹뉴스
많이 본 뉴스
주간 클릭
주간 댓글 >
주간 검색어
집계 안내 >


주간 댓글


< 2018.01.06 - 01.12. > 

종합 정치 **경제** 사회 생활/문화 세계 IT/과학 연예 포토 TV

1 

[종합]법무장관 "가상화폐는 도박...거래 금지 특별법 추진"
【과천=뉴시스】 박주성 기자 = 박상기 법무부 장관이 11일 오전 정부과천청사 법무부 3층 브리핑실에서 법조기자단...  뉴시스 | 2018-01-11

2 

법무장관 "가상화폐는 도박...거래소 폐쇄 정부 법안 준비"
【과천=뉴시스】 박주성 기자 = 박상기 법무부 장관이 11일 오전 정부과천청사 법무부 3층 브리핑실에서 법조기자단...  뉴시스 | 2018-01-11

데이터 수집

수집한 댓글 개수

2,316,305

- ▶ 기간 : 17.09.30 ~ 17.12.08 (10주)
- ▶ 일주일에 카테고리당 30개 기사의 댓글 크롤링

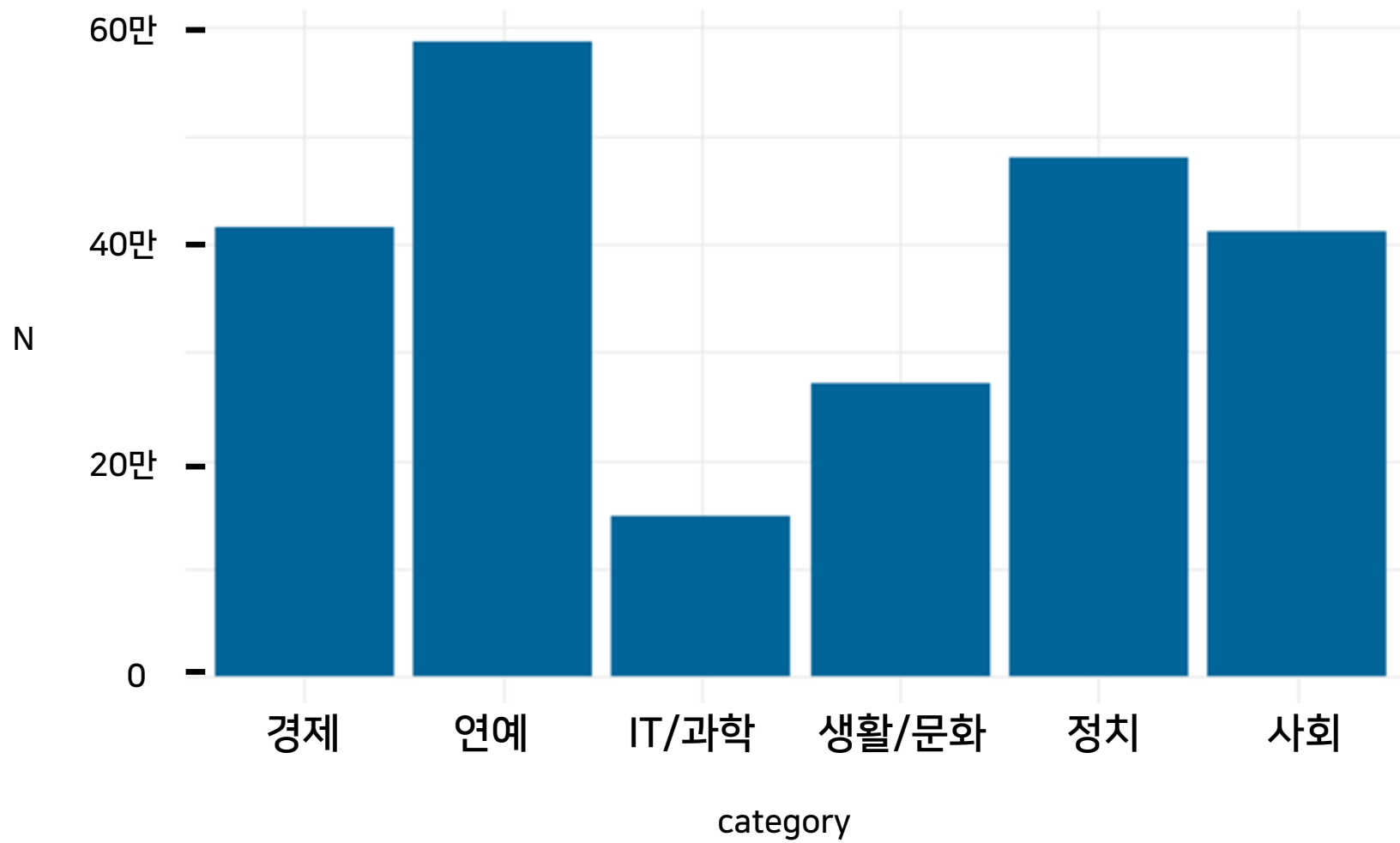
수집한 카테고리 개수

6

- ▶ 사회, 경제, 생활/문화, IT/과학, 정치, 연예면

데이터 수집

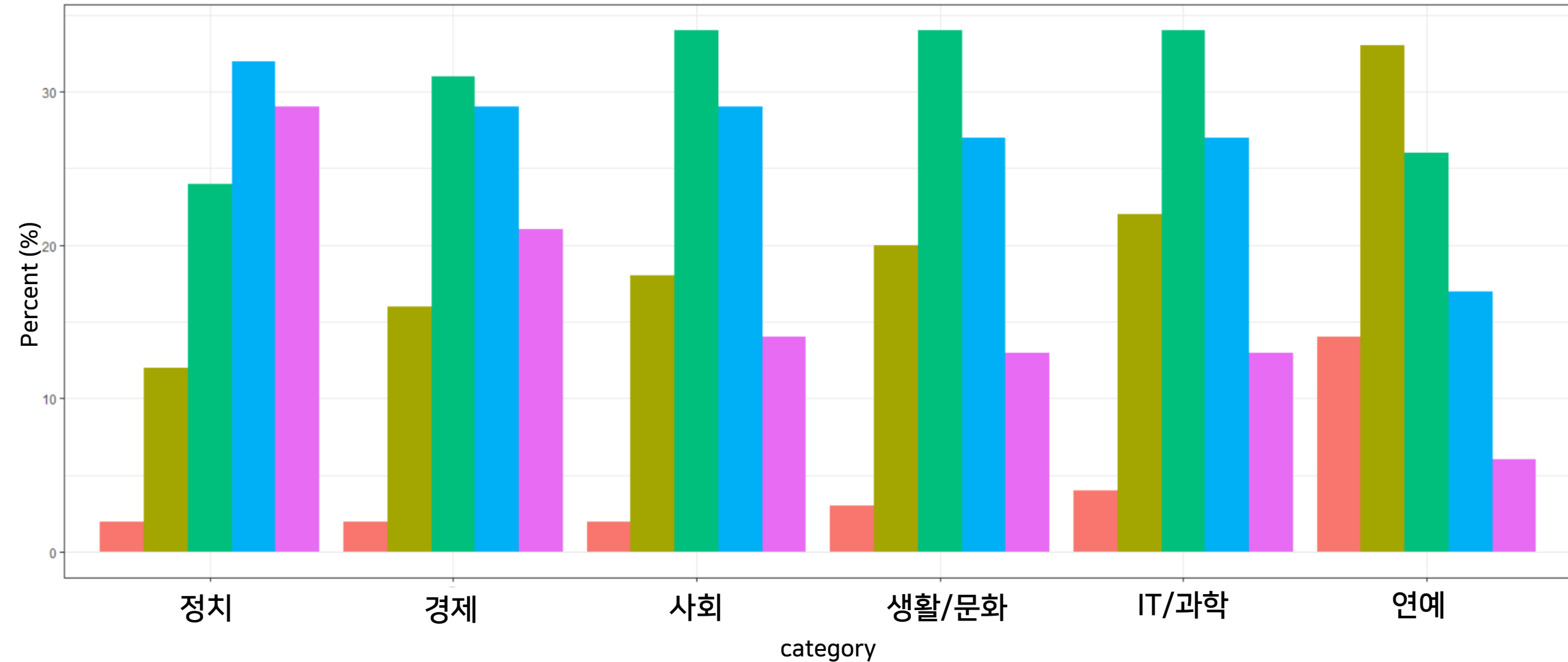
- 카테고리별 댓글 개수



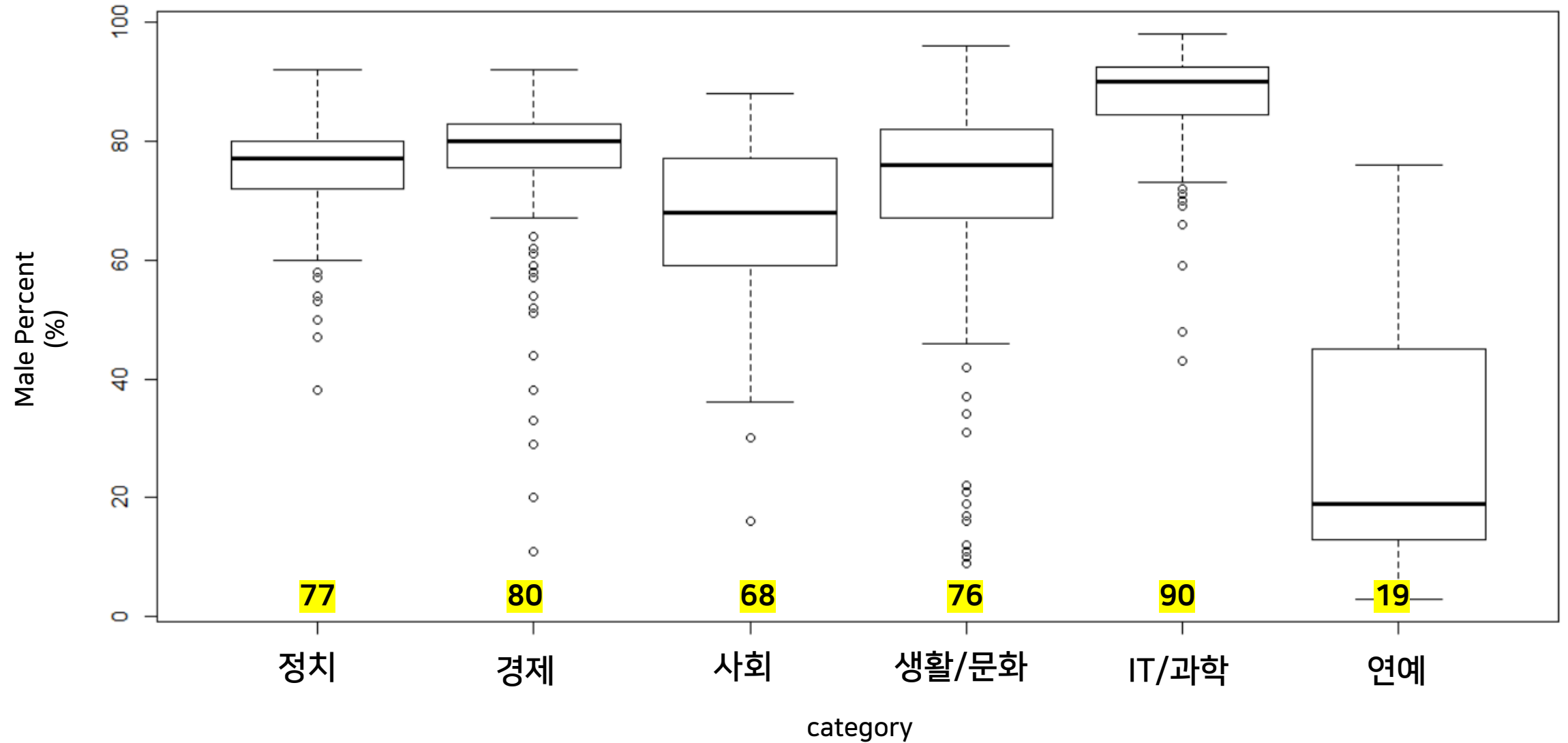
데이터 수집

- 카테고리별 연령 비율

age
10대 20대 30대 40대 50대 ↑



- 카테고리별 남자 비율



악성 댓글 정의

인터넷상에서 상대방이 올린 글에 대해 **비방**이나 **혐담**을 하는 **악의적인 댓글**.

악성 댓글은 사이버 범죄이자 언어폭력으로, **근거를 갖춘 부정적 평가와 구별**해야 함.

비속어 정의

'격이 낮고 속된 말'을 뜻하나, 그 기준이 달리 정해져 있지는 않음.

넓은 의미의 비속어로 일상생활에서 상대를 무시 · 모욕하거나 불쾌감 등을 주는 말들

• 비속어 기준

비속어 ○

뒤진다	짱깨
돌았다	조센징
지랄	대가리
년	아가리
존나	주둥이
~충	빨갱이
한남	ㅈㄹㅇ
메갈 관련 용어	쓰레기
일베 관련 용어	씨부리다
	시부리다
	뺨다
	인간말종

비속어 △

놈
개~
개돼지
쳐~
레밍
정몽주니어
기레기
언레기
~빨다

비속어 X

멍청하다	국개의원
짓거리	개한민국
꼬라지	견찰
저 따위	개검
관종	~빠
종자	빠순이
헬조선	호구
사이비	병맛
좀비	열라
거지	문재앙
~꼴	문죄인
개뿔	닭그네
꼴통	쥐박이
북괴	
경북괴	

• 비속어 기준

비속어 ○

뒤진다	짱깨
돌았다	조센징
지랄	대가리
년	아가리
존나	주둥이
~충	빨갱이
한남	ㅈㄹㅇ
메갈 관련 용어	쓰레기
일베 관련 용어	씨부리다
	시부리다
	뺨다
	인간말종

비속어 △

놈
개~
개돼지
쳐~
레밍
정몽주니어
기레기
언레기
~빨다

비속어 X

멍청하다	국개의원
짓거리	개한민국
꼬라지	견찰
저 따위	개검
관중	~빠
종자	빠순이
헬조선	호구
사이비	병맛
좀비	열라
거지	문재앙
~꼴	문죄인
개뿔	닭그네
꼴통	쥐박이
북괴	
경북괴	

- 비속어 기준

비속어 ○

지랄
~충
쓰레기

비속어 △

놈
개돼지
기레기

비속어 X

멍청하다
관종
거지

- 악성 라벨링을 위한 샘플링

- 한 개의 뉴스당 4개의 댓글 층화 추출
- 총 7,180개 댓글 샘플

악성 지수 구축

- 악성 댓글 기준

	0 (악플 아님)	1 (악플 가능성)	2 (악플)
비속어	-	○	○
공격성	-	-	○
표현의 자유	○	○	-

▶ 팀원 다섯명 라벨의 최빈값을 최종 라벨로 사용

악성 지수 구축

- 악성 댓글 기준

	0 (악플 아님)	1 (악플 가능성)	2 (악플)
비속어	-	○	○
공격성	-	-	○
표현의 자유	○	○	-

▶ 공격성 < 표현의 자유

악성 지수 구축

- 악성 댓글 기준

	0 (악플 아님)	1 (악플 가능성)	2 (악플)
비속어	-	○	○
공격성	-	-	○
표현의 자유	○	○	-

▶ 공격성 > 표현의 자유

- 악성 지수 라벨링 예시

- 0 “추석 연휴 방역에 힘 써주시는 분들 감사하고 조심하세요 ㅠㅠ”
- 1 “저 정도면 영아살인미수죄에 해당된다. 그런데?? 6개월이라....지랄 꿀갑도 참 풍년이다...”
- 2 “에라이~~미친년도 급수가 있으면 넌 최고레벨감이다.”

'twitter-Korean' 을 이용한 형태소 분석

- 트위터에서 만든 한국어 형태소 분석기
- 트위터 게시글과 같은 짧은 글에 특화되어 있어 댓글 데이터에 적합함
- 정규화, 토큰화, 어근화, 어구추출 수행
- 의미를 담고 있는 일부 품사만 선택

명사(Noun), 동사(Verb), 형용사(Adjective), ㄱ (Korean Particle), 숫자(0~9), 알파벳(a,b,...)



안녕하세요 우리는 매의 눈 입니다ㅋㅋㅋ



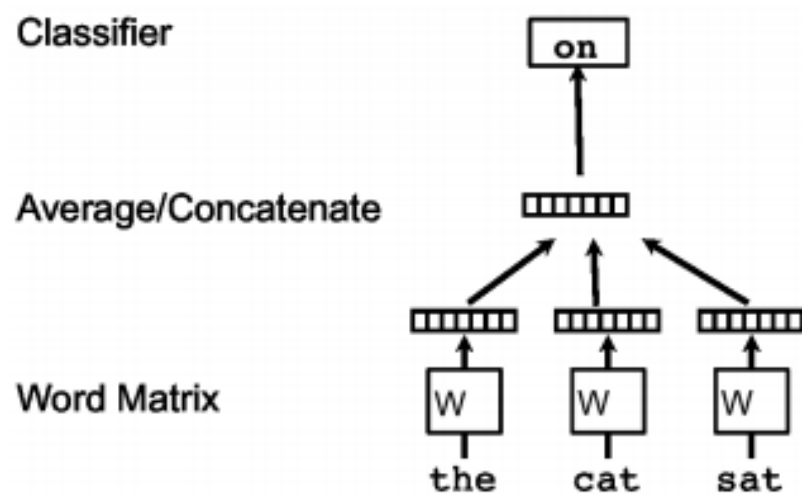
안녕하다/Adjective, 우리/Noun, 는/Josa, 매/Noun, 의/Josa, 눈/Noun, 이다/Adjective, ㅋㅋ/KoreanParticle



모델 구현

Word2vec

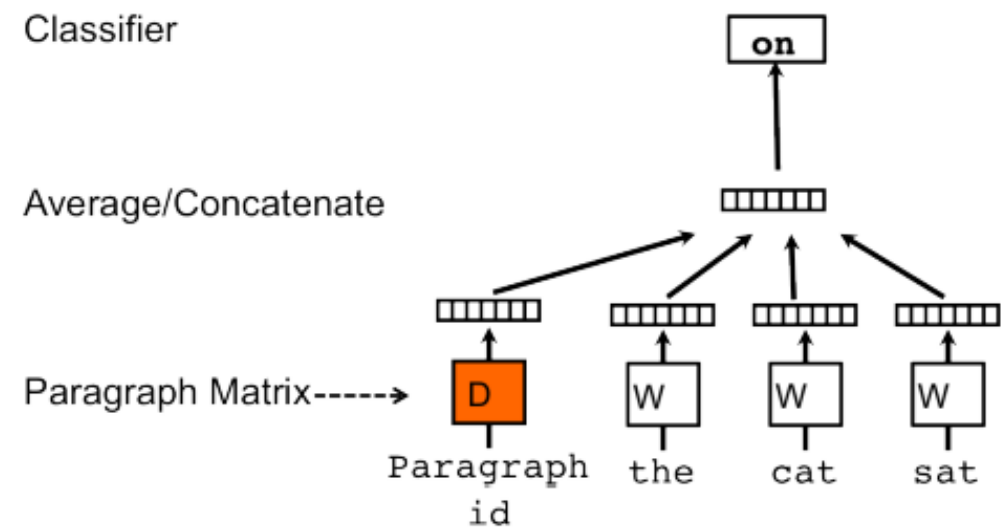
단어의 의미 자체를 벡터화.
문맥으로부터 단어를 예측



<CBOW>

Doc2vec

단어벡터와 같은 공간에
문장 또는 문서를 벡터화.



<PV-DM>

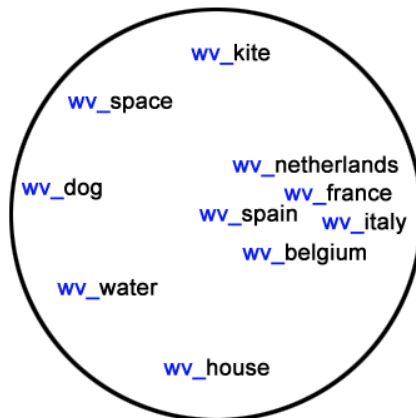
분석 기법

word2vec

Input:
text

Lorem ipsum dolor
sit amet, conse-
tur adipiscing elit,
sed diam nonumy
eiusmod tempor
invidunt ut labore
et dolore magna
aliquam erat, sed
diam voluptua. At
vero eos et

Model:



train for
each word
a word vector

vector space:
consists of **word vectors**
for each word

most_similar('france'):

spain	0.678515
belgium	0.665923
netherlands	0.652428
italy	0.633130

highest cosine
distance values
in vector space
of the nearest
words

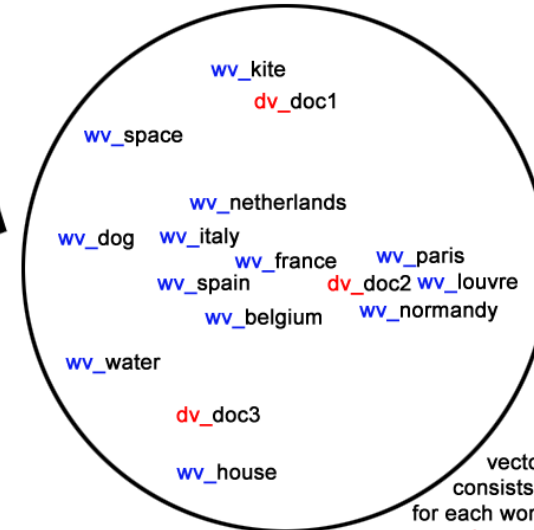
doc2vec

Input:
many document

Lorem ipsum dolor
sit amet, conse-
tur adipiscing elit,
sed diam nonumy
eiusmod tempor
invidunt ut labore
et dolore magna
aliquam erat, sed
diam voluptua. At
vero eos et

doc1,
doc2,
doc3 ...

Model:



training a word
vector for each
word and each
document gets
an ID/tag with
a vector while
training

most_similar('france'):

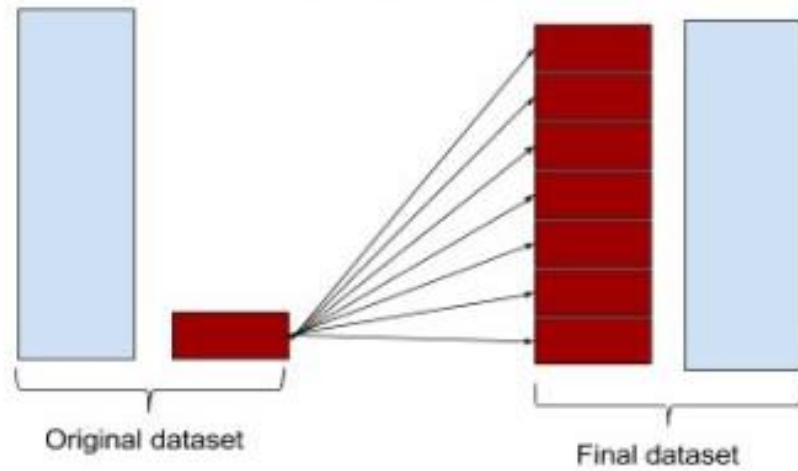
paris	0.876543
louvre	0.765432
normandy	0.654321
...	

highest cosine
distance values
in vector space
with consideration
of the document
vectors

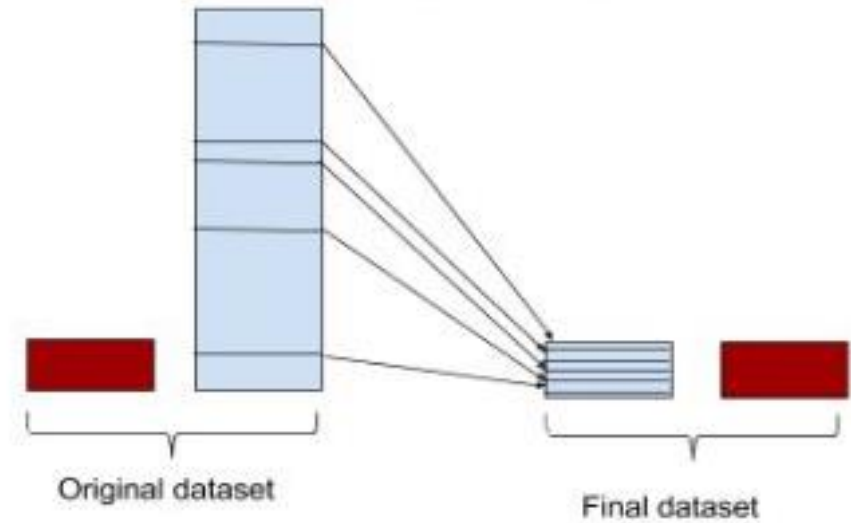
vector space:
consists of **word vectors**
for each word and additional
document vectors

- 클래스 불균형 해소

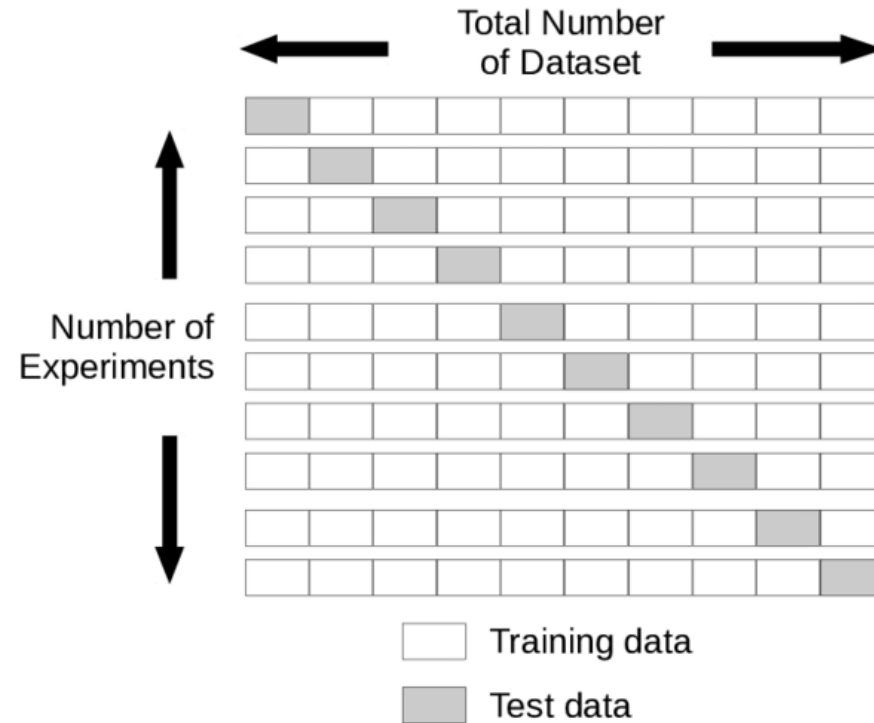
Up sampling



Down sampling



- 10-fold Cross Validation



- 최적의 하이퍼 파라미터를 구하기 위한 모델 튜닝에 사용 → 일반화 성능 만족
- 층화 10-fold CV → 카테고리별, 라벨별로 층화 추출

- **Hyper parameter tuning optimization**

- **Manual Search**

- 값을 직접 대입하며 찾아간다.

- **Grid Search**

- 범위를 지정하고 값을 대입하며 찾아간다.

- **Random Search**

- 범위를 지정하고 랜덤하게 값을 찾아간다.

- 일정한 시간 안에 결과를 내야하는 경우 좋은 결과를 내는 것으로 알려졌다.

- **Bayesian Optimization**

- 지금까지의 실험 결과를 바탕으로 통계 모형을 만들어 값을 탐색한다.

- **모델 선정**

	AdaBoost	Neural Network	Random Forest	SVM
mean_score	0.849	0.831	0.873	0.880

→ 성능이 가장 좋은 SVM 선택

- 모델 성능 평가 : 예측 정확도

		TRUE Label		
		0	1	2
PREDICTED Label	0	603	47	23
	1	9	9	1
	2	3	2	20

▶ Baseline $\left(\frac{603+9+3}{717}\right) * 100 = 85.8\%$

▶ Accuracy $\left(\frac{603+9+20}{717}\right) * 100 = 88.15\%$

▶ Type I error $\left(\frac{9+3+2}{717}\right) = \frac{14}{717}$

▶ Type II error $\left(\frac{47+23+1}{717}\right) = \frac{71}{717}$

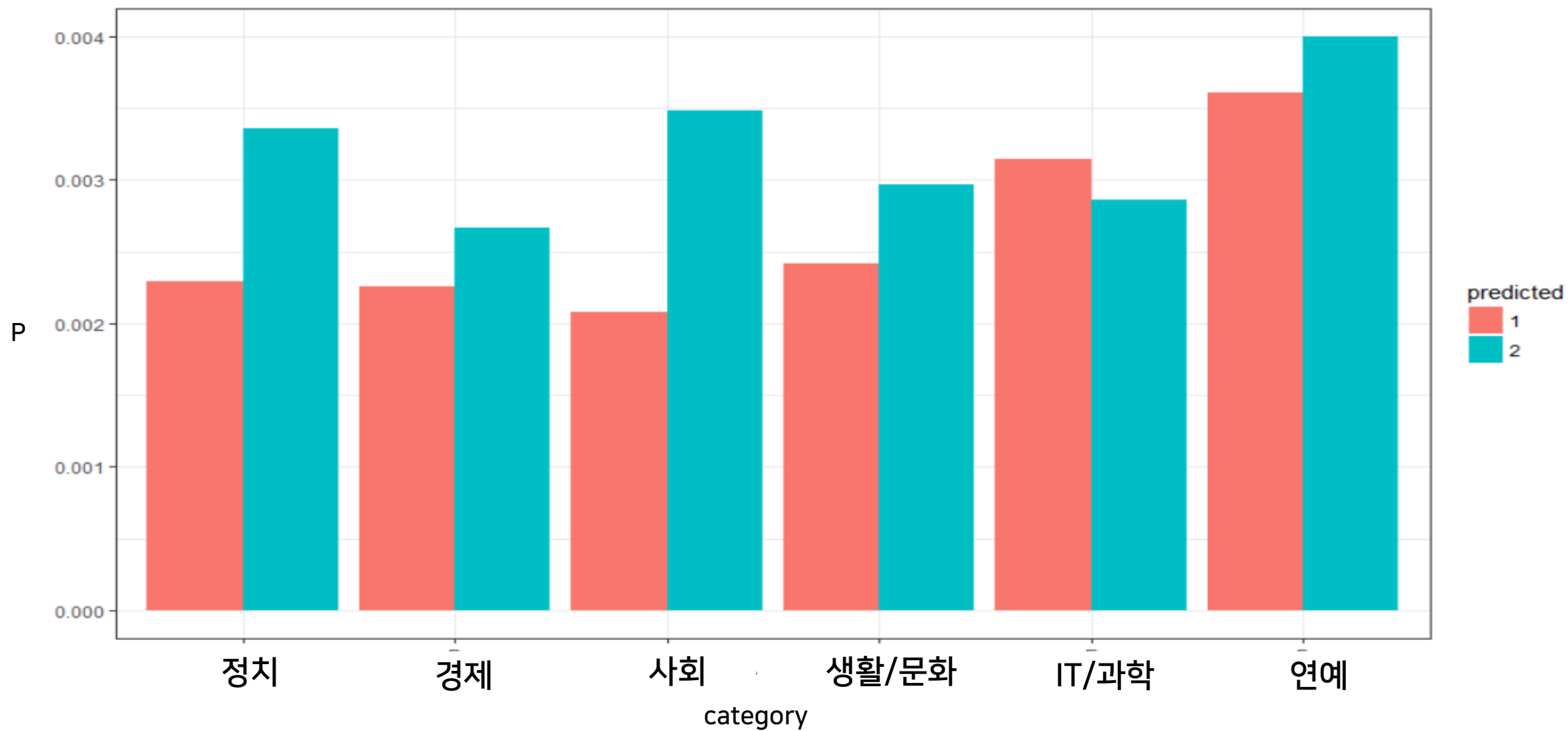
IV

결론

1. 전체 네이버 기사 댓글 EDA

2. 네이버 기사에서 벗어난 범용성 있는 데이터 적용 결과

- 카테고리별 악성 지수 비율



- 카테고리별 악성 댓글 wordcloud

- 정치



- 악성 댓글이 많은 기사 Title

- 문 대통령, 내일 하루 휴가..."연말에 남은 연차휴가 다 소진"
- 김종대 "저는 이국종 교수를 겨냥하지 않았습니다"
- 文대통령, 추석구상 양대 화두 '북한·협치'
- 노무현·이명박·박근혜... 동시에 칼 겨눈 검찰
- 홍준표 "권양숙 여사, 노 전 대통령 가족들 고발 검토"

- 경제



- 1인당 월 13만 원 최저임금 지원"...세금 퍼주기?
- 트럼프, 한미 FTA '미치광이 전략'...오늘 밤 협상 변수
- [앵커&리포트] 광군제 위력에 韓 패션·화장품 특수... '유커'들도
- 반복되는 한화 총수 일가의 '갑질'...3남 김동선씨 또 만취난동
- '엄마 돈' 10억원으로 강남 아파트 산 의사...지능탈세 백태

- 사회



- “맘카페에 찍히면 망해요” 동네 식당 · 카페 주인 속앓이
- 집에서 기르던 개한테 물려 1세 여아 숨져
- [미집행 20년, 사형제를 말하다] '1분이 1년' 같았던 그날의 형장... '평생의 회한'에 울었다
- [현장영상] 이영학 "꿈만 같이 느껴져...죄송하다"
- 野 “김이수 대행체제는 위헌”... 헌재國監 파행 끝 연기

- 생활/문화



- '10만발의 불꽃' 여의도 불꽃축제, 최고 명당은 어디?
- 경주 동궁 우물서 나온 1천년 전 인골 얼굴 공개된다(종합)
- 부산 가을바다 수만발 불꽃
- "아직 50분인데..." 알람 울리기 전 저절로 눈 떠지는 이유는?
- [너도나도 케이블카] ① "대박 사업" vs "환경 적폐"

- IT/과학



- 네이버, 프로축구연맹 청탁에 불리한 기사 재배열...대표 사과
- 완전'자급제'는 물 건너갔다...자급제 비중 확대 초점
- 中우주정거장 지구 추락 위기...충돌 위치 미궁
- [IF] 우주서 오래 머물면 뇌 형태 바뀐다
- 文 정권 방통위, 종편 재승인 심사 엄격하게 한다

- 연예



- 워너원 박지훈 中팬들, 결식아동돕기 동참...모금액 184만원 돌파 '훈훈 팬心'
- "내마음속에저장♥" 네티즌 선정 2017 인터넷 유행어 1위
- "귀엽거나, 섹시하거나"...강다니엘 화보 촬영 모습 공개
- [포토]워너원 박지훈 '이 비주얼 실화?'
- [포토]꽃보다 '라이관린'

• 새로운 테스트 셋

You **Tube** Live



실시간 채팅

- 보임
- 낮에뜨는별 홈 이더가 홍콩은 상승인데 여긴 왜이래
- T TigerOJ 해외오르면 또 사람들 득달같이와요 ㅎ
- 선차이 이더 갓뜨아~!
- 니가가라하와이 한국 거래소 내에 거래량이 없으니 깐
- 낮에뜨는별 줍줍..... 단타 치실분들 참고
- 니가가라하와이 거래소에서 변동폭이 없는거임
- 너는내 달빛 님 거래량 더줄면 끝났다고바야지ㅋㅋ
ㅋㅋ 나머지 거래량은 ㅋㅋ그냥 시체들이 준버하는거고 ㅋ



Ryungpue @jk4beer

#아미는_방탄소년단의_얼굴이다

1일 전 | 리트윗 0 | 관심글 0

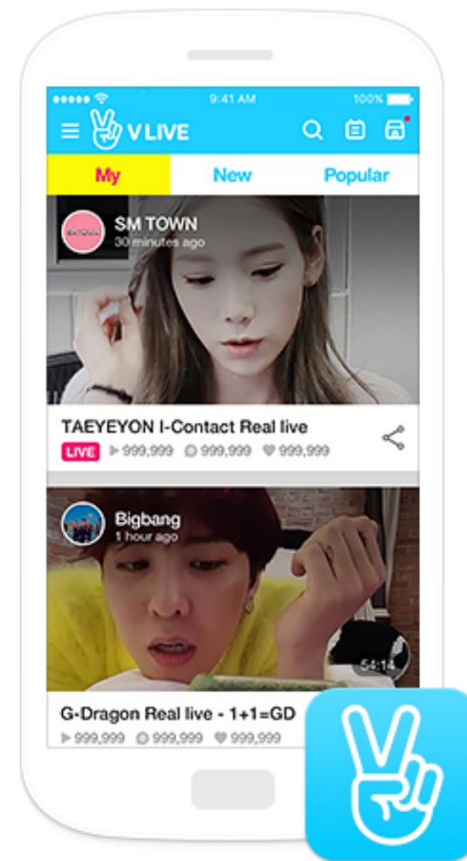


이소민

지민이 멍 때리는것 조차 귀여워♡
7분 전



Run BTS! 2018 - EP.35
vlive.tv



• 예측 정확도

		TRUE Label		
		0	1	2
PREDICTED Label	0	127	45	12
	1	0	6	0
	2	0	2	8

▶ Baseline

$$\left(\frac{127+0+0}{200}\right) * 100 = 63.5\%$$

▶ Accuracy

$$\left(\frac{127+6+8}{200}\right) * 100 = 70.5\%$$

▶ Type I error

$$\left(\frac{0+0+2}{200}\right) = \frac{2}{200}$$

▶ Type II error

$$\left(\frac{45+12+0}{200}\right) = \frac{57}{200}$$

- **개선할 점**

각 댓글에 대한 성별, 연령 등 개인정보를 알 수 없음

- ▶ 깊이 있는 EDA가 어려웠음

악성 라벨링 과정에서 주관성을 배제할 수 없음

- ▶ 많은 사람들의 라벨링 표본으로 객관성 확보

감성 분석을 수행하면 더 높은 모델 성능을 기대할 수 있음

매의 눈 시연

QnA

THANK YOU!