워드 임베딩 기법을 이용한 혐오성 어휘집합의 자동 확장

2022. 11. 18.

조단비, 강승식
신한은행 디지털혁신단, 국민대학교 인공지능학부

목차

서론

1) “**익명 제도**” 악용

> 개인 정보 보호 / 표현의 자유

> 욕설, 비난, 차별 등의 혐오 표현 증가 / 사회적 문제 초래

2) 딥러닝 모델 활용한 NLP 시스템의 “**윤리적 문제**”

Ex1) 번역 시스템

“걔는 간호사야” => “She’s a nurse” / “걔는 경찰이야” => “He’s a cop”

Ex2) ‘이루다’ 챗봇 시스템

“출산은 징그러워”, “소름끼친다”, “전라도는 싫어”

혐오 표현 탐지 연구

관련 연구

> Waseem (2016)

- 언어 체계 문제에 대처하기 위해 음절 n-gram 토큰 자질 추출
- CNN+GRU 모델 적용(79%)

> Moon (2020)

- 한국어 뉴스 악성 댓글 데이터셋 구축 (label: none, offensive, hate)
- 현재 공개된 유일한 한국어 혐오 발언 데이터셋

> 신민기 (2021)

- 심심이 데이터셋 활용 및 KoELECTRA 모델 적용 (66.1%)
- 나쁜말 점수 기반으로 라벨링 작업 수행 (나쁜말 점수: 수동)

워드 임베딩 기반의 혐오성 어휘사전 구축

혐오

위키백과, 우리 모두의 백과사전.

혐오(嫌惡)는 어떠한 것을 증오, 불결함 등의 이유로 싫어하거나 기피하는 감정으로, 불쾌, 기피함, 싫어함 등의 감정이 복합적으로 이루어진 비교적 강한 감정(사람이 느끼는 것을 기준으로 함)을 의미한다.

> Keywords: 혐오 / 증오 / 불결 / 불쾌 / 기피

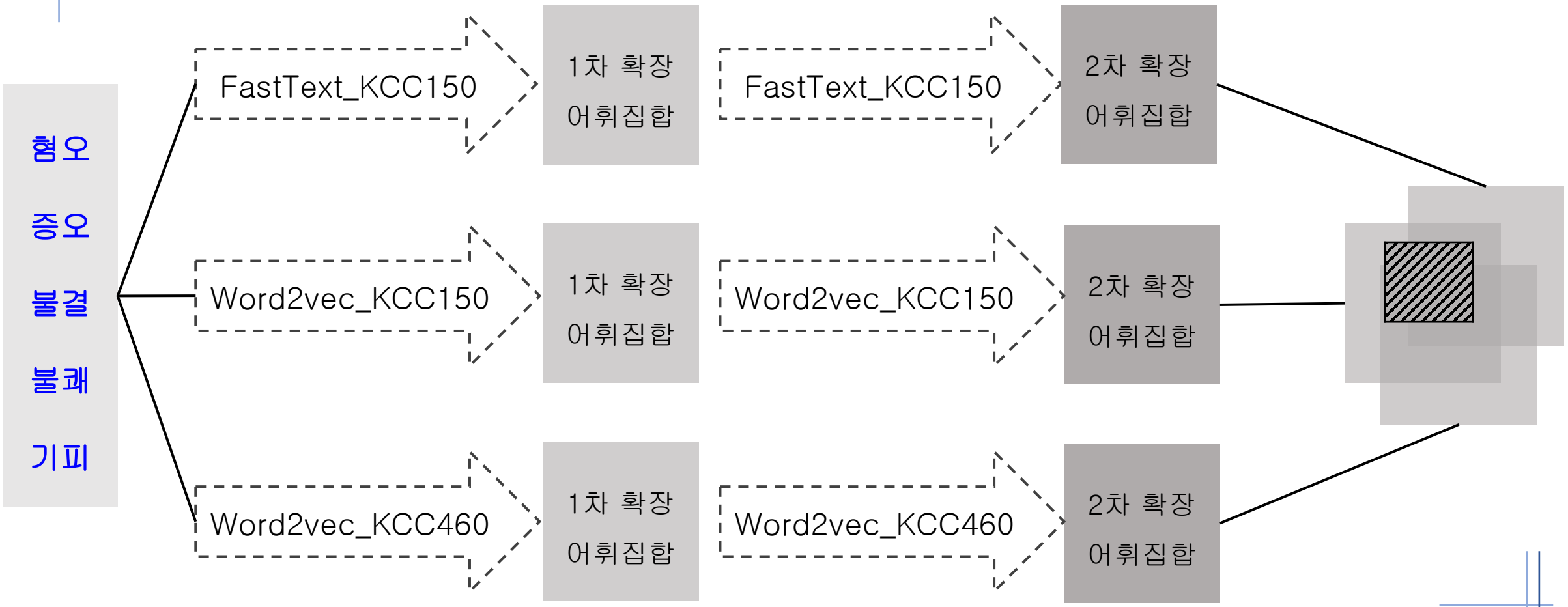
> 사전학습한 벡터 모델 (vector size: 300)

모델	학습 말뭉치	사전 크기
FastText_KCC150	KCC150	546,686
Word2vec_KCC150	KCC150	546,686
Word2vec_KCC460	KCC460	1,078,554

> 말뭉치 정보

말뭉치	문장 수	어절 수
KCC150	11,961,347	150,705,457
KCC460	29,316,426	467,649,207

워드 임베딩 기반의 혐오성 어휘사전 구축



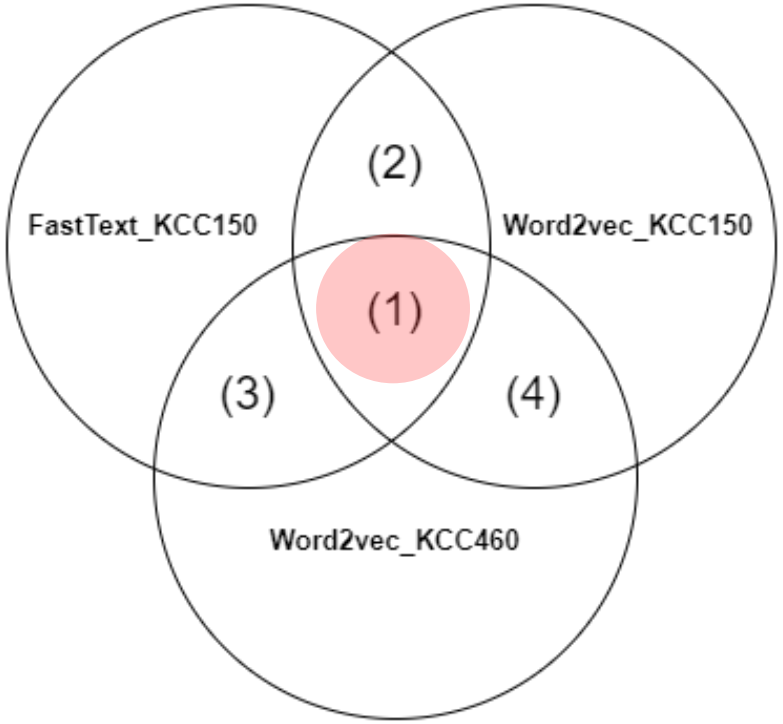
어휘집합 구성: 1차 확장

	FastText_KCC150	Word2vec_KCC150	Word2vec_KCC460
혐오	24 (증오 / 혐오감 / 경멸 / 혐오주의 / 폭력적)	190 (증오 / 경멸 / 혐오감 / 폭력적 / 천박)	124 (여성혐오 / 증오 / 인종주의 / 혐오발언 / 혐오표현)
증오	44 (증오심 / 분노 / 적개심 / 혐오 / 경멸)	326 (경멸 / 적개심 / 증오심 / 복수심 / 미움)	114 (혐오 / 증오심 / 인종주의 / 적개심 / 반유대주의)
불결	5 (더러운 / 비위생적 / 더럽고 / 지저분한 / 지저분하고)	16 (비위생적 / 더러운 / 더럽고 / 지저분한 / 청결)	8 (비위생적 / 지저분하고 / 더럽고 / 지저분한 / 청결)
불쾌	27 (불쾌감 / 무례 / 당혹 / 당황 / 서운한)	127 (무례 / 언짢은 / 당황 / 불쾌감 / 서운한)	43 (불쾌감 / 무례 / 모욕적 / 당혹 / 언짢은)
기피	2 (꺼리는 / 꺼려하는)	1 (꺼리는)	2 (꺼리는 / 꺼려하는)
총합	102	660	291

어휘집합 구성: 2차 확장

	FastText_KCC150	Word2vec_KCC150	Word2vec_KCC460
혐오	1,492 (703) (공포감 / 고정관념 / 우월주의자 / 백인우월주의 / 두려움)	41,000 (5,522) (비판 / 박탈감 / 관능적 / 질타 / 고립감)	9,482 (2,355) (반무슬림 / 반난민 / 이슬람교도 / 페미니즘 / 소수인종)
증오	1,256 (489) (열만 / 잔혹 / 절망 / 질투 / 좌절감)	81,230 (9,481) (무참히 / 앙갚음 / 기괴 / 핍박 / 자기)	12,066 (2,956) (유태인 / 이데올로기 / 반대세력 / 인종차별 / 반공)
불결	94 (67) (비위생 / 너저분하고 / 지저분하던 / 지저분하다 / 지저분하면)	4,656 (2,722) (이로운 / 추잡 / 분진 / 위생적 / 해롭다는)	189 (128) (위생적 / 위생 / 청결히 / 썩는 / 더러워진)
불쾌	1,056 (739) (고통스러운 / 거북스러움 / 서운한 / 마땅하다는 / 어이없다는)	46,811 (10,095) (서운함 / 싫다는 / 싫었던 / 배신감 / 싫고)	3,715 (1,798) (서운함 / 모독 / 난감 / 곤혹 / 고통스러운)
기피	32 (26) (꺼려하는 / 꺼리는 / 싫어하는 / 꺼리고 / 꺼린다는)	11 (11) (꺼려하는 / 기피 / 꺼리고 / 싫어하는 꺼려)	48 (33) (꺼려하는 / 꺼리는 / 꺼려지는 / 부담스러워 / 꺼린다)
총합	3,930	173,708	25,500

혐오성 어휘집합 자동구축



	사전 크기
FastText_KCC150	1,687
Word2vec_KCC150	16,530
Word2vec_KCC460	4,904
(1)	717
(2)	279
(3)	77
(4)	2,355

> 혐오 어휘 예시
증오 / 혐오 / 경멸 / 자학 / 광기

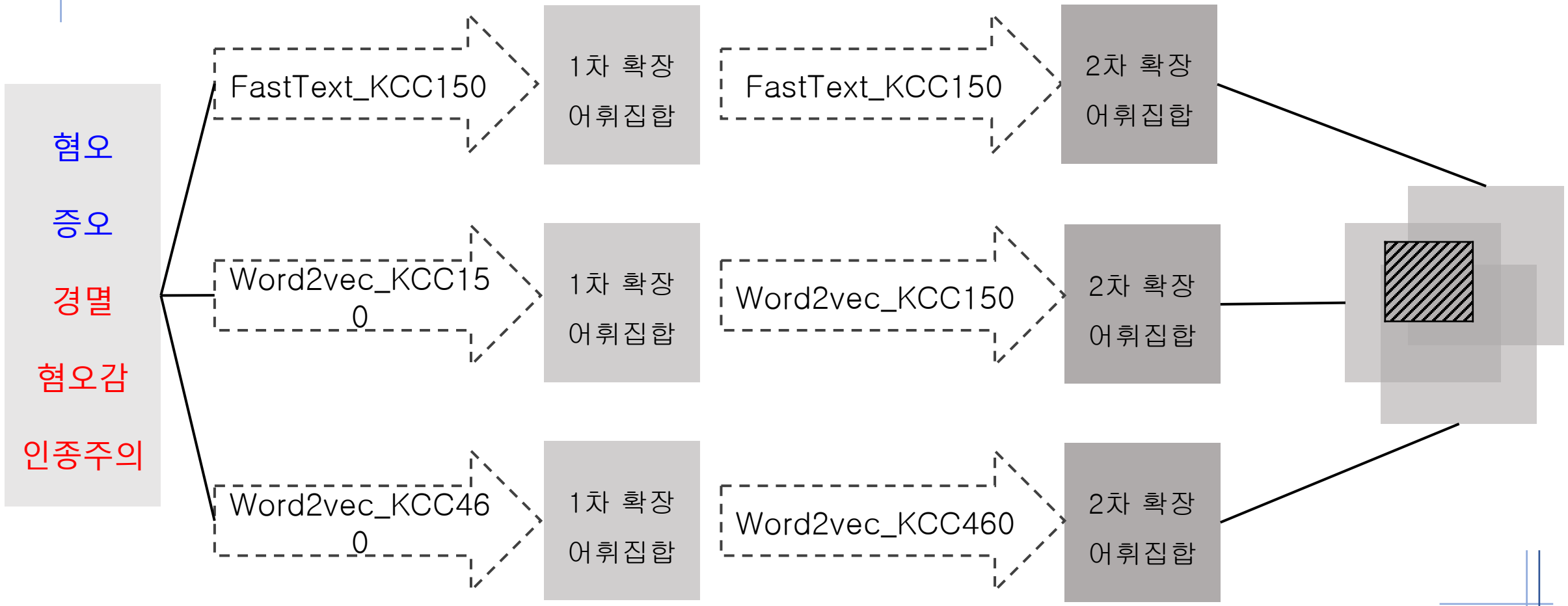
혐오성 키워드 집합: '혐오'와 유사도 높은 어휘

벡터 모델	증오	경멸	혐오감	인종주의
FastText_KCC150	0.67	0.62	0.66	0.54
Word2vec_KCC150	0.75	0.74	0.69	0.60
Word2vec_KCC460	0.78	0.65	0.58	0.73
평균	0.73	0.67	0.64	0.62

핵심 키워드 seed 집합

- 1안: { 혐오, 증오, 불결, 불쾌, 기피 }
- 2안: { 혐오, 증오, 경멸, 혐오감, 인종주의 }

워드 임베딩 기반의 혐오성 어휘사전 구축



어휘집합 구성: 1차 확장

키워드별 1차 확장 어휘수

	FastText_KCC150	Word2vec_KCC150	Word2vec_KCC460
혐오	203	931	580
증오	14	207	92
경멸	1	260	10
혐오감	4	67	3
인종주의	0	7	114
총합	222	1,472	799

Word2vec_KCC150을 이용한 1차 확장 예시

Word2vec_KCC150		
혐오	증오	
증오: 0.76 경멸: 0.74 혐오감: 0.70 폭력적: 0.66 천박: 0.65	경멸: 0.59 적개심: 0.59 증오심: 0.58 복수심: 0.58 미움: 0.57	
경멸	혐오감	인종주의
증오: 0.58 멸시: 0.58 모멸: 0.55 혐오: 0.55 냉소: 0.54	불쾌감: 0.53 모멸감: 0.52 공포감: 0.52 적대감: 0.52 동정심: 0.51	국수주의: 0.45 파시즘: 0.45 식민주의: 0.43 국가주의: 0.42 민족주의: 0.42

어휘집합 구성: 2차 확장

키워드별 2차 확장 어휘수

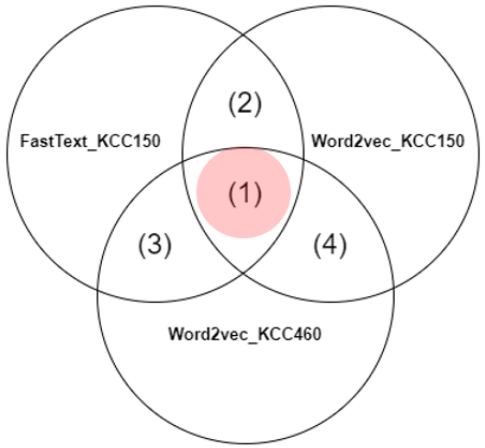
	FastText_KCC150	Word2vec_KCC150	Word2vec_KCC460
혐오	907 (460)	30,827 (4,027)	8,584 (2,154)
증오	10 (6)	4,537 (853)	671 (327)
경멸	0 (0)	1,408 (261)	36 (29)
혐오감	3 (3)	117 (68)	5 (5)
인종주의	0 (0)	8,584 (2154)	1,469 (404)
총합	920	44,065	10,765

Word2vec_KCC460을 이용한 2차 확장 예시

Word2vec_KCC460		
혐오		증오
고정관념: 0.45 혐오범죄: 0.43 괄시: 0.43 선입관: 0.43 Speech: 0.43		적대감: 0.45 불신감: 0.39 김치녀: 0.39 적대심: 0.38 복수심: 0.38
경멸	혐오감	인종주의
냉대: 0.37 모멸: 0.35 핍박: 0.35 질시: 0.35 여성혐오: 0.34	공포감: 0.35 굴욕감: 0.33 모욕감: 0.32 성적수치심: 0.31 모멸감: 0.30	증오: 0.42 인종차별적: 0.41 인종차별주의자:0.40 인종차별: 0.38 인종혐오: 0.37

임베딩 모델별 어휘집합의 크기

사전학습 벡터 모델	어휘 개수	혐오성 어휘 예시
FastText_KCC150	460	증오: 3.75 두려움: 3.42 적대감: 3.06 열등감: 2.67 죄의식: 2.65
Word2vec_KCC150	4,039	경멸: 74.62 증오: 67.70 죄의식: 66.06 열등감: 63.30 증오심: 57.77
Word2vec_KCC460	2,164	인종주의: 33.03 혐오: 28.82 국수주의: 27.60 반유대주의: 24.04 극우주의: 19.76



교집합을 통한 혐오 어휘사전 구축

교집합	사전 크기	혐오 어휘 예시
(1)	261	경멸: 89.78 증오: 88.76 혐오: 75.86 열등감: 73.22 죄의식: 71.57
(2)	44	부끄러움: 51.69 자책감: 43.21 외로움: 27.09 노여움: 26.65 정열: 15.76
(3)	39	반유대주의: 24.65 백인우월주의: 18.28 인종차별주의: 14.23 인종주의자: 13.50 성차별주의: 12.24
(4)	808	자학: 60.16 천박: 56.56 나약: 45.66 야비: 41.10 교만: 39.89

혐오성 추정 및 시각화

문장 $S = (t_1, t_2, \ldots, t_n)$

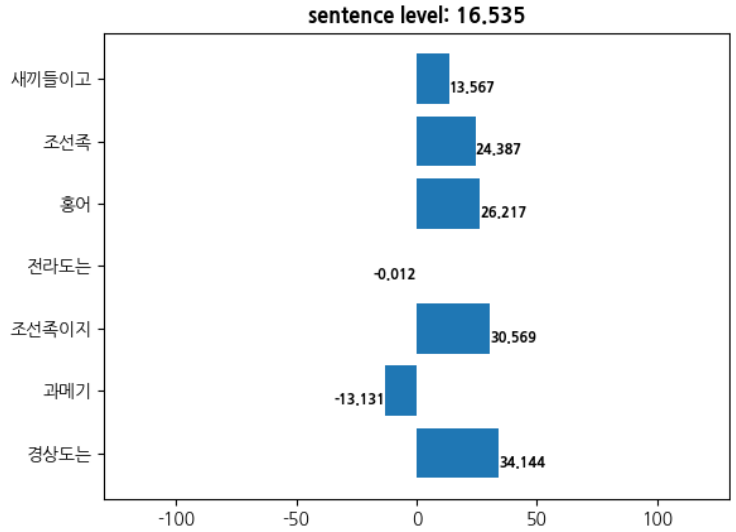
HS_t: 토큰 t의 혐오성

HS_S: 문장 S의 혐오성

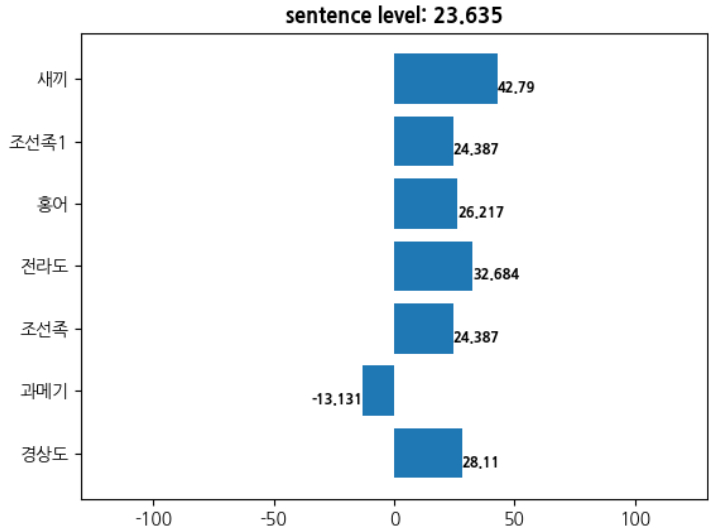
$$HS_t = \sum_{i=1}^{len(D)} sim(t, k_i), \ k_i \in D$$

$$HS_S = \frac{1}{n} \sum_{j=1}^{n} HS_{t_j}$$

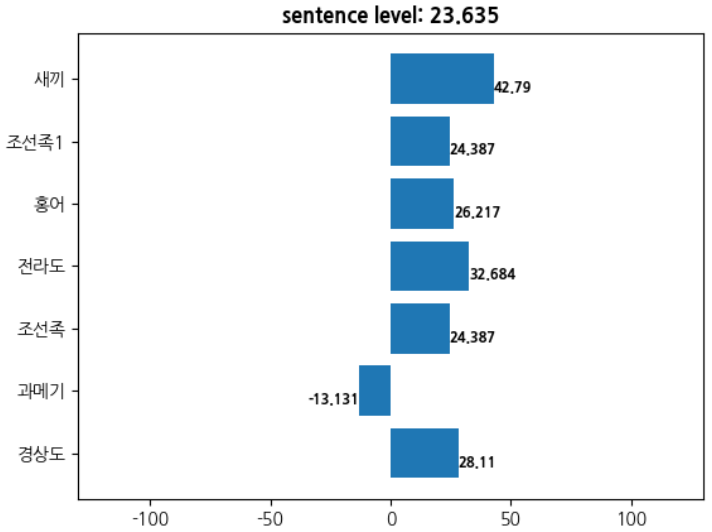
예시) “경상도는 과메기 조선족이지 전라도는 홍어 조선족 새끼들이고“



어절



Okt 명사



KLT2000 명사

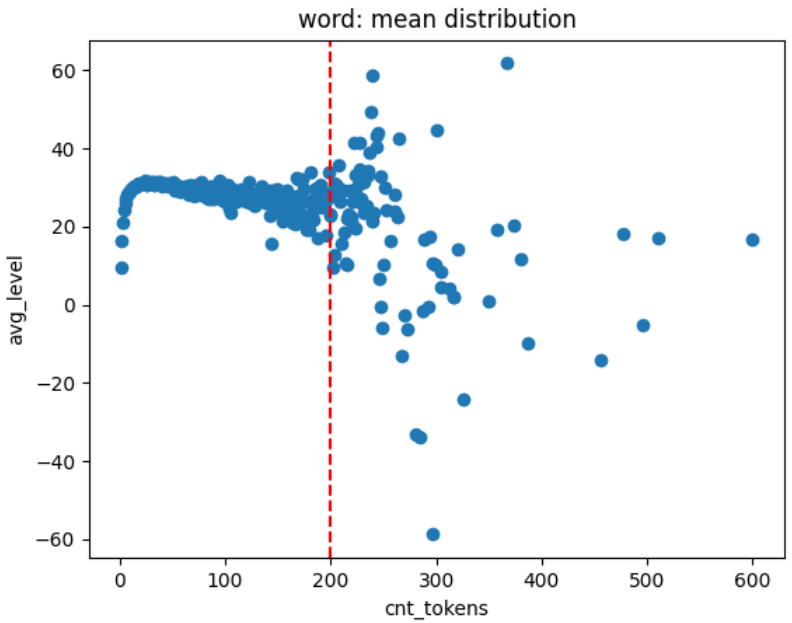
데이터

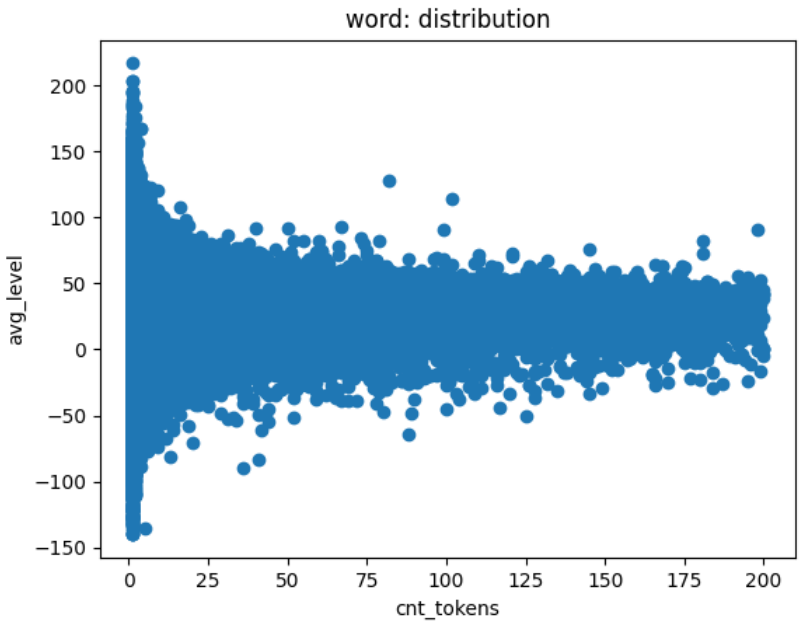
> 수집 데이터: 일베 댓글 데이터셋

- 댓글 수: 1,388,264개 (중복제거: 1,196,345개)

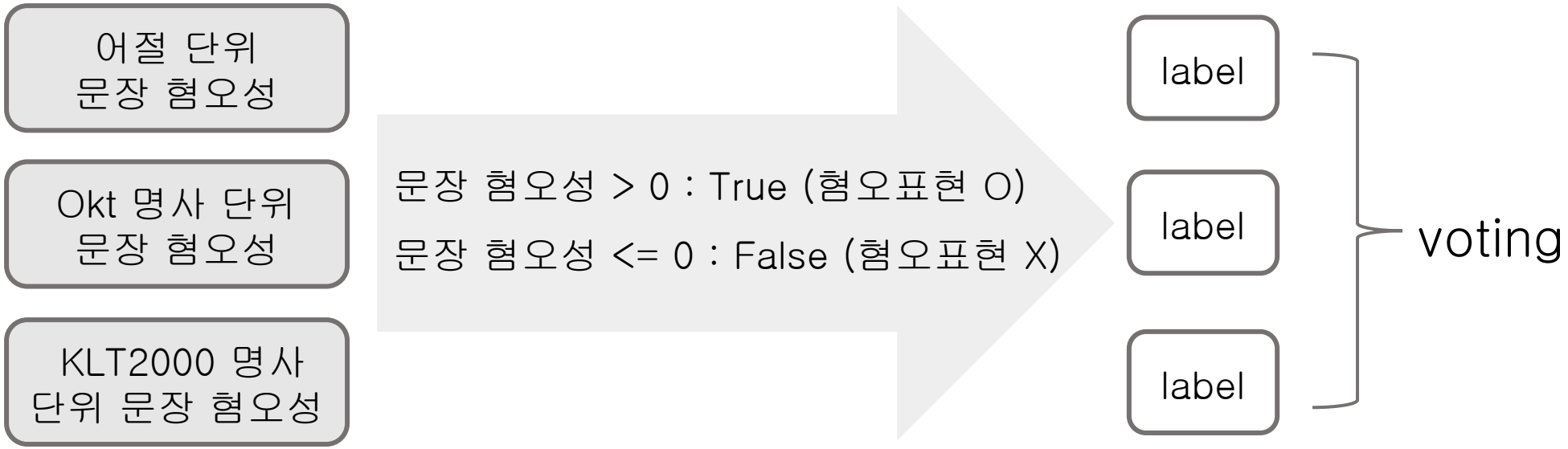
> 전처리 작업

- 토큰 수 200 초과 데이터 삭제 (동일한 토큰 반복 및 특수문자) : 1,196,017





데이터 라벨링



> 라벨링 결과

True: 974,191 & False: 221,826

> 데이터 균형을 위한 다운샘플링 수행

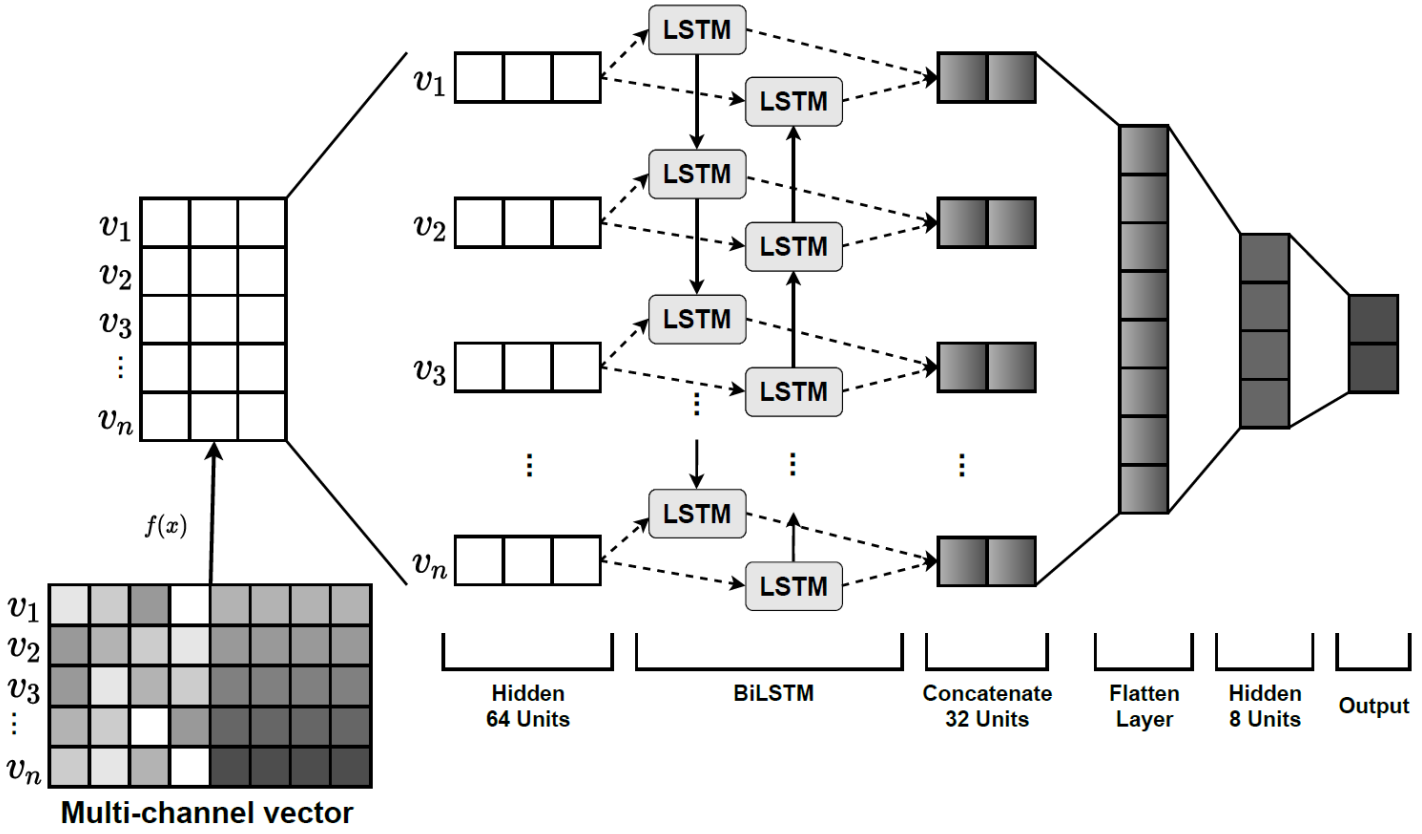
True: 221,826 & False: 221,826 = total 443,652

> 학습 데이터 분할 8:2

	문장 수	어절 수	Okt 명사 수	KLT2000 명사 수
훈련 데이터	354,920	2,029347	1,760,928	1,621,570
평가 데이터	88,732	481,711	424,047	385,889
총합	443,652	2,511,058	2,184,975	2,007,459

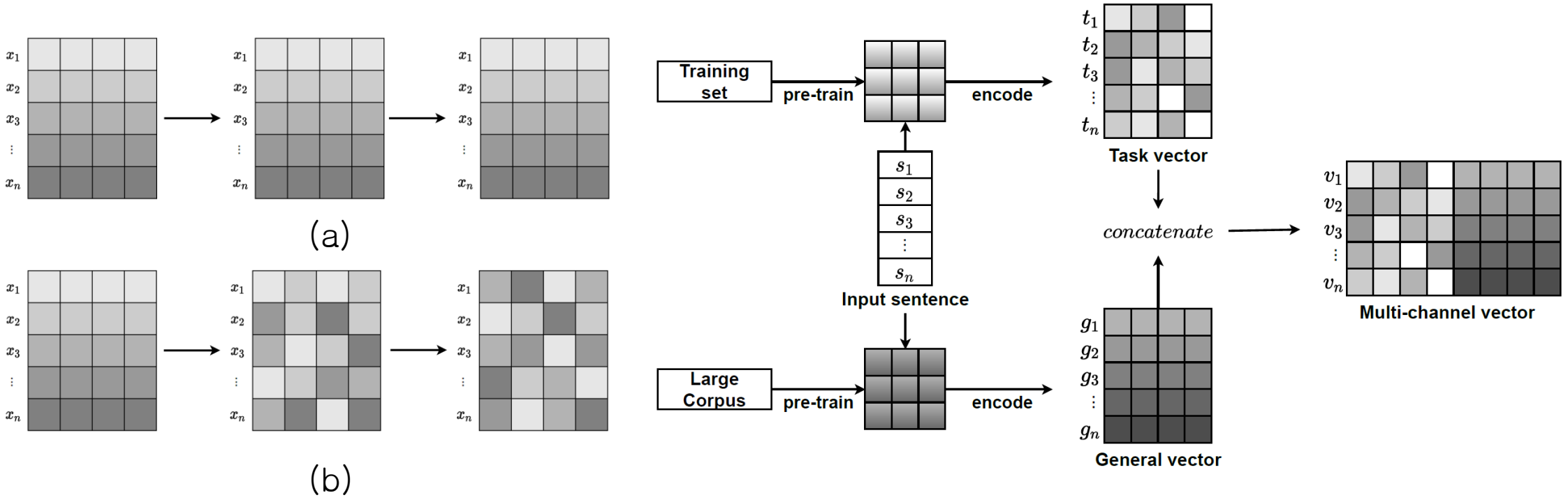
딥러닝을 이용한 혐오 표현 탐지

> Multi-channel LSTM



Parameter	Value
Epoch	25
Batch size	32
Learning rate	0.01
Optimizer	SGD

Multi-channel LSTM



실험결과 비교

핵심 키워드 seed 집합

- 1안: { 혐오, 증오, 불결, 불쾌, 기피 }
- 2안: { 혐오, 증오, 경멸, 혐오감, 인종주의 }

실험 및 결과: Baseline (LSTM)

{ 혐오, 증오, 불결, 불쾌, 기피 }

Embedding model	Input token	ACC(%)	F1 (%)	FNR
FastText_KCC150	Word	81.78	80.73	0.24
	Okt	80.92	79.79	0.25
	KLT2000	83.22	82.26	0.22
Word2vec_KCC150	Word	81.01	79.86	0.25
	Okt	81.28	81.74	0.18
	KLT2000	82.66	81.61	0.23

{ 혐오, 증오, 경멸, 혐오감, 인종주의 }

Embedding model	Input token	ACC(%)	F1(%)	FNR
FastText_KCC150	Word	83.32	83.23	0.17
	Okt	82.31	82.30	0.18
	KLT2000	83.48	83.34	0.17
Word2vec_KCC150	Word	80.79	80.77	0.19
	Okt	82.33	82.95	0.14
	KLT2000	82.80	82.34	0.19

실험 및 결과: Multi-channel LSTM

{ 혐오, 증오, 불결, 불쾌, 기피 }

Vector model	Input token	ACC(%)	F1-score(%)	FNR
FastText_KCC150	Word	84.00	82.73	0.23
	Okt	83.61	83.33	0.18
	KLT2000	86.71	85.91	0.19
Word2vec_KCC150	Word	81.89	81.04	0.23
	Okt	82.68	82.46	0.19
	KLT2000	83.69	82.78	0.22
Word2vec_KCC460	Word	82.27	81.77	0.20
	Okt	82.43	82.60	0.17
	KLT2000	83.49	82.98	0.20

{ 혐오, 증오, 경멸, 혐오감, 인종주의 }

Vector model	Input token	ACC(%)	F1(%)	FNR
FastText_KCC150	Word	85.16	84.22	0.21
	Okt	83.87	84.41	0.13
	KLT2000	86.85	86.25	0.18
Word2vec_KCC150	Word	81.88	82.11	0.17
	Okt	82.94	82.70	0.18
	KLT2000	83.81	83.23	0.19
Word2vec_KCC460	Word	82.03	81.77	0.19
	Okt	82.74	82.56	0.18
	KLT2000	83.78	82.87	0.21

결과 분석

> 실제 혐오 표현을 일반 표현으로 예측한 오류 유형

(1) 자음 표현 (ex. ㅇㅂ, ㅅㅂ 등)

(2) 적은 어휘 수 (ex. 토끼래 *발 ㅋㅋ)

(3) 변형된 언어 (ex. * 같은 십*까 아닥해)

> 실제 일반 표현을 혐오 표현으로 예측한 오류 유형

(1) ‘ㅋ’이 많은 경우 (ex. ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 적절한 콩트다)

(2) 욕설 포함 (ex. *나 잘만들었넼ㅋ)

(3) 잘못된 라벨링 (ex. *주작 ㅁㅈㅎ)

혐오 표현의 일부 *로 표시

결론

> 일베 댓글 데이터셋 구축

: 혐오 표현에 대한 제재 당위성 부여 및 사회적 문제 해결을 위한 혐오 표현 탐지

> 혐오성 어휘집합의 자동 확장

: 워드 임베딩 기법을 이용한 1차 확장 및 2차 확장

> Multi-channel LSTM을 이용한 혐오 표현 탐지 실험

: 정확도 86.711% & F1-score 85.914%

> 혐오표현 탐지 오류의 유형 분석

감사합니다.