Improving Generalizability in Implicitly Abusive Language Detection with Concept Activation Vectors

저자: Isar Nejadgholi, Kathleen C. Fraser, Svetlana Kiritchenko
발제자: 박채원
22-11-18

Abstract

- 끊임없이 변화하는 실제 적용 데이터에 대한 기계 학습 모델의 Robustness는 매우 중요
- 특히 인간의 웰빙에 영향을 미치는 문제에 대해선 더욱이 중요시 되기 때문에 욕설(abusive) 탐지 시스템은 정확성을 유지하기 위해 정기적으로 업데이트 되어야 함

- 일반적인 욕설 분류기가 도메인 외의 명시적(explicit) 욕설 감지는 신뢰할 수 있는 결과를 내지만 암시적(implicit) 욕설 감지엔 어려움이 있다는 것을 보임
- 컴퓨터 비전의 TCAV 방법을 기반으로 해석가능성 기술을 제안
  - 명시적, 암시적 욕설에 대한 모델의 민감도를 정량화
  - 또한 이를 사용하여 새로운 개념의 데이터에 대한 모델의 일반화 가능성을 설명
- 새로운 명확성(explicit) 측정 방법론 도입
- 암시적(implicit) 데이터를 이용한 증강을 통해 모델 강화 가능

Introduction

- 세가지 질문
  - 개념(concept)을 어떻게 텍스트로 공식화하는가?
  - 새로운 개념이 등장할 때 훈련된 분류기의 민감도를 어떻게 정량화(수치화) 하는가?
  - 분류기를 어떻게 신뢰할 수 있도록 업데이트 하는가?
- 소셜 미디어에서의 코로나 관련 아시아인 인종차별
  - 코로나가 예상치 못하게 발생함으로써 새로운 용어가 나타나고, 아시아 커뮤니티를 향한 증오 발언 강조됨
  - 이러한 새로운 이벤트가 특정 유형의 혐오 발언을 강화할 수는 있지만 문제의 근본인 경우는 드물다.
  - 즉 그러한 혐오는 사건 이전에도 존재하기 때문에 분류기가 이러한 표현을 아예 못잡지는 않는다.
- 이 연구에서 중요한 포인트
  - 텍스트가 명시적(직접적) 또는 암시적(간접적) 혐오를 표현하는지 여부
  - 암시적 혐오를 이해하기 위해서는 코로나와 같은 상황에 대한 사전 지식이 필요로 됨
  - 그렇기 때문에 사전 훈련된 분류기가 이러한 데이터를 다루기는 특히 어려움
- TCAV(Testing Concept Activation Vector)
  - 분류기가 **특정 개념**을 **클래스**와 연관 시키는지 여부를 설명하는 데 사용된다
  - 즉, 코로나 관련 아시안 차별 개념을 정의하고 분류기가 이를 positive(abusive)와 연관시키는지 묻는다.

Introduction

- 데이터 증강
  - 원본 데이터셋이 아닌 다른 소스의 데이터를 추가하여 훈련 데이터를 강화하는 과정을 지칭
  - 인간이 정의한 concept에 대한 민감도를 이용해 데이터 증강이 가능한가
  - 분류기를 업데이트할 때 분류기가 아직 민감하지 않은 개념의 데이터를 추가하는 데 중점을 두어야 함
- 기존 능동 학습 프레임워크 (Conventional active learning framework)
  - 분류 신뢰도가 낮은 데이터를 가장 유익한, 정보가 많은 데이터라고 제안
  - 그러나 신뢰할 수 있는 불확실성 추정치를 제공하지 못하는 심층 신경망의 무능력은 이 방법을 선택하길 어렵게 함
- 이 논문에서는 implicit한 데이터가 분류기를 업데이트하는 데 가장 유익하다고 제안
  - 하지만 명확성 정도를 측정할 수 있는 정량적 방법이 없기 때문에 명확성 수치화를 위해 TCAV를 제안하고 이를 효율적인 데이터 증강에 사용
- 기여
  - TCAV 프레임워크를 구현해 분류기의 민감도를 정량화
  - 분류기가 **명시적**인 코로나 관련 아시안 차별에 대해서는 잘 일반화(분류) 되지만, **암시적인** 인종 차별에 대해서는 일반화되지 않음을 보임
  - TCAV 방법을 통해 데이터 증강을 하는 방법 제안

Dataset

- 4개의 영어 데이터셋
  - Founta, wiki -> 코로나 이전 구축된 데이터
  - EA, CH -> 코로나 이후 구축된 데이터 (코로나 관련 아시안 인종차별을 대상으로 함)
- 이진화
  - 모든 데이터셋을 positive(abusive, hate)와 negative(other)로 이진화함

Dataset	**Data Source**	**Positive Class**	**Negative Class**	**Number (%Pos:%Neg)**		
				Train	*Dev*	*Test*
Wikipedia Toxicity (*Wiki*) (Wulczyn et al., 2017)	Wikipedia comments	Toxic	Normal	43,737 (17:83)	32,128 (9:91)	31,866 (9:91)
Founta et al. (2018) dataset (*Founta*)	Twitter posts	Abusive; Hateful	Normal	62,103 (37:63)	10,970 (37:63)	12,893 (37:63)
East-Asian Prejudice (*EA*) (Vidgen et al., 2020)	Twitter posts	Hostility against an East Asian entity	Criticism of an East Asian entity; Counter speech; Discussion of East Asian prejudice; Non-related	16,000 (19:81)	1,200 (19:81)	2,800 (19:81)
COVID-HATE (*CH*) (Ziems et al., 2020)	Twitter posts	Anti-Asian COVID-19 hate; Hate directed to non-Asians	Pro-Asian COVID-19 counterhate; Hate-neutral	–	–	2,319 (43:57)

Dataset

- 어휘의 차이
  - 새로운 주제가 등장함에 따라 어휘가 바뀜
  - Founta와 wiki는 코로나 이전에 수집되었기 때문에 chinavirus, wuhanflu와 같은 단어를 포함하지 않으며, 관련 단어(ex| 'china')의 빈도가 이후의 데이터와 다를 수 있다.
- 이를 확인하기 위해 각 데이터의 positive class(hate, abusive)에서 가장 빈번하게 사용되는 100개의 단어 추출
  - 각 데이터셋 간 겹침을 계산

1) 코로나 관련 단어 (COVID-related)
2) 일반적으로 비속하고 혐오스러운 단어 (Hateful)
3) 기타 모든 단어 (Other)

- 코로나 이후 데이터끼리, 코로나 이전 데이터끼리 더 많은 단어를 공유

Datasets	Count	Shared Words
EA - CH	50	**COVID-related (32%):** ccp, 19, communist, pandemic, coronavirus, covid19, chinesevirus, infected, covid, chinese, chinavirus, corona, wuhanvirus, wuhan, china, virus **Hateful (0%)** **Other (68%):** racist, came, want, country, calling, come, does, spread, like, amp, media, eating, did, human, world, know, government, say, started, think, need, blame, evil, time, people, don, new, let, news, stop, countries, just, spreading, make
Wiki - Founta	37	**COVID-related (0%)** **Hateful (30%):** *ss, b*tch, id*ot, n*ggas, d*ck, f*cking, f*ck, sh*t, hell, hate, stupid **Other (70%):** oh, dont, want, way, going, come, does, like, look, life, did eat, sex, know, say, think, man, need, time, people, said, stop, really, just, make, tell
Founta - EA	19	**COVID-related (0%)** **Hateful (0%)** **Other (100%):** racist, want, calling, come, does, like, did, world, know, say, think, need, time, people, trying, let, stop, just, make
Wiki - EA	15	**COVID-related (0%)** **Hateful (0%)** **Other (100%):** people, want, did, say, think, good, need, come, does, stop, just, know, like, make, time
Founta - CH	35	**COVID-related (0%)** **Hateful (23%):** *ss, b*tch, f*cking, f*ck, sh*t, hate, stupid, f*cked **Other (77%):** racist, want, way, going, calling, come, does, like, got, look, did, eat, world, know, say, think, man, trump, need, time, people, said, let, stop, really, just, make
Wiki - CH	33	**COVID-related (0%)** **Hateful (27%):** *ss, b*tch, f*cking, f*ck, sh*t, hate, stupid, shut, kill **Other (73%):** want, way, going, come, does, like, look, did, eat, right, know, die, say, think, man, need, time, people, don, said, stop, really, just, make

Dataset

- CH와 일반 데이터가 EA와 일반 데이터보다 더 많은 단어 공유
  - CH에 더 많은 명시적 욕설이 포함되어 있을 것이다.

- 명시적 정도에 대한 라벨링
  - CH와 EA의 positive 데이터에 규칙을 사용해 추가적으로 라벨링 진행
    - 코로나에 대한 지식 없이도 욕설로 식별될 수 있는 데이터는 명시적
    - 나머지는 암시적으로 주석
  - CH는 85%가 명시적이지만 EA는 8%만이 명시적으로 분류됨
  - CH와 EA가 코로나 관련 어휘를 공유하지만 명시성 측면에서는 매우 다름
    - CH – 명시적 데이터를 많이 포함
    - EA – 암시적 데이터를 많이 포함

Cross-Dataset Generalization

- 새로운 도메인의 욕설 데이터를 통해 사전 훈련된 분류기의 견고성을 평가
- 새로운 도메인의 암시적 및 명시적 혐오에 대한 분류기의 일반화 가능성에 대해 어휘 변경의 영향을 관찰
- Wiki-exp는 explicit general abuse와 randomly sampled negative 데이터로 학습된 분류기
- Wiki와 founta의 train 데이터의 pos-neg 비율이 다르지만 성능은 비슷
- CH는 데이터 크기가 작지만 성능이 잘 나옴
- 혐오 탐지의 교차 데이터셋 일반화가 종종 클래스 크기 보다는 훈련 및 테스트 레이블이 정의 및 샘플링 전략의 호환성에 의해 좌우 된다고 주장

Domain	Train Set	AUC		F1-score	
		EA	*CH*	*EA*	*CH*
COVID	*EA*	0.94	0.82	0.74	0.66
	CH	0.86	-	0.62	-
pre-COVID	*Founta*	0.69	0.73	0.29	0.65
	Wiki	0.64	0.74	0.27	0.69
	Wiki-exp	0.58	0.71	0.15	0.56

- 일반 분류기들이 CH(명시적)에는 비교적 잘 수행되지만 EA에 대해서는 어려움을 겪음
- 따라서 일반 분류기는 새 도메인의 **암시적 남용**을 학습하도록 업데이트 되어야 함

Sensitivity to Implicit & Explicit abusive to Explain Generalizability

- 일반화는 암시적 및 명시적 혐오에 대해 별도로 평가되어야 함
  - 하지만 별도의 암시적 및 명시적 테스트셋을 구축하는 데는 너무 많은 비용이 듬
- TCAV 제안
  - 적은 수의 데이터만을 이용해 명시적 및 암시적 코로나 관련 인종차별에 대한 분류기의 민감도 계산
  - 사용자가 선택한 개념이 학습 중에 특성으로 직접 사용되지 않았더라도 예측에 얼마나 중요한지를 측정하기 위한 학습 후 해석 방법이다.
  - 개념은 개념의 예시들로 정의된다
    - Ex) zebra 클래스와 관련된 시각적 개념으로 stripe를 제안, stripe가 포함된 이미지를 수집해 stripe개념을 정의
- 언어 기반 TCAV
  - 개념(concept) -> 수동으로 개념 예시들을 선택되고 이에 개념 주석을 달아줌
  - 이렇게 선택된 예시들의 표현을 평균화하여 개념을 벡터로 나타냄
  - 이를 통해 개념 활성화 벡터(CAV)는 분류기의 활성화 공간에서 개념을 표현하도록 학습됨
  - 그리고 방향 도함수를 사용해, 개념의 방향을 향한 입력 변화에 대한 예측 민감도를 계산
  - 개념의 방향을 바꾸는 입력의 예측 민감도

Sensitivity to Implicit & Explicit abusive to Explain Generalizability

- RoBERTa 기반 분류기에 TCAV 절차 적용
  - K개의 단어, n차원 입력공간
  - 입력 텍스트 $x \in \mathbb{R}^{k \times n}$
  - RoBERTa 인코더를 다음과 같이 고려 $f_{emb} : \mathbb{R}^{k \times n} \rightarrow \mathbb{R}^m$
  - 입력 텍스트를 RoBERTa 표현([CLS] 토큰에 대한 표현)에 매핑 $r \in \mathbb{R}^m$
  - 각 컨셉 C 마다 예시 N_c개를 수집하고 이를 RoBERTa 표현에 매핑 $r_C^j, j = 1, ..., N_C$
  - 랜덤하게 선택된 컨셉 예시 N_v개의 RoBERTa 표현을 평균내서 CAV v_C^p 를 계산 ($N_v < N_c$)

$$v_C^p = \frac{1}{N_v} \sum_{j=1}^{N_v} r_C^j \quad p = 1, .., P \qquad (1)$$

  - v_C^p (개념)에 대한 Positive class의 개념적 민감도는 도함수로 계산 가능하다
    - h는 RoBERTa의 표현을 positive 클래스의 logit값으로 매핑하는 함수

$$S_{C,p}(x) = \lim_{\epsilon \to 0} \frac{h(f_{emb}(x) + \epsilon v_C^p) - h(f_{emb}(x))}{\epsilon}$$

$$= \nabla h(f_{emb}(x)).v_C^p \qquad (2)$$

Sensitivity to Implicit & Explicit abusive to Explain Generalizability

- 수학식 2에서 $S_{C,p}(x)$는 클래스 로짓의 변화를 측정한다.
- 일련의 입력 X에 대해서 개념 C 방향의 작은 변화에 대한 입력(X)의 비율로 TCAV score를 계산

$$TCAV_{C,p} = \frac{|x \in X : S_{C,p}(x) > 0|}{|X|} \qquad (3)$$

- 1에 가까운 TCAV 점수는 대부분의 입력 예에서 로짓값이 증가함을 나타냄
- 식 3은 개념에 대한 점수 분포를 정의
- 개념에 대한 분류기의 전반적인 민감도를 결정하기 위해 이 분포의 평균 및 표준 편차를 계산

Classifier's Sensitivity to a Concept

- Nc=100의 Concept C를 정의하고 6개의 개념으로 실험
  - 베이스라인을 위해 일관성 없는 개념을 형성하기 위한 랜덤 트윗
  - 코로나 관련 키워드를 포함하는 무작위 트윗을 사용해 혐오스럽지 않은 코로나 관련 트윗
  - EA의 dev와 CH에서 명시적인 욕설로 라벨링 된 트윗 선택
  - EA에서의 암시적인 아시안 차별 개념
  - CH에서의 암시적인 아시안 차별 개념
    - 두개의 서로 다른 데이터셋에서 예제를 선택하는 것이 민감도에 영향을 미치는지 여부를 평가
  - Founta dev set에서 일반적인 혐오 발언 트윗

Non-coherent concept: random tweets collected with stop words as queries
COVID-19: tweets collected with words *covid, corona, covid-19, pandemic* as query words
Explicit anti-Asian abuse: tweets labeled as explicit from *EA* dev and *CH*
Implicit abuse (EA): tweets labeled as implicit from *EA* dev
Implicit abuse (CH): tweets labeled as implicit from *CH*
Generic hate: tweets from the *Hateful* class of *Founta* dev

Classifier's Sensitivity to a Concept

- 실험
  - 각 개념에 대해 P=1000 CAV를 계산
  - CAV는 랜덤하게 선택된 개념 예시 Nv=5개의 평균이다.
  - 2000개의 무작위 트윗을 입력 예제로 사용
  - 각 데이터셋으로 학습된 분류기에 대한 positive class의 각 개념 TCAV 점수 분포의 평균과 표준 편차를 나타냄

Classifier	Concept					
	non-coherent	COVID-19	explicit anti-Asian	implicit (EA)	implicit (CH)	generic hate
EA	0.00 (0.00)	0.00 (0.00)	**0.90** (0.26)	**0.87** (0.30)	**0.70** (0.42)	0.00 (0.00)
CH	0.00 (0.00)	0.00 (0.00)	-	0.35 (0.44)	-	0.21 (0.12)
Founta	0.00 (0.02)	0.00 (0.01)	**0.92** (0.22)	0.00 (0.06)	0.19 (0.32)	**0.60** (0.44)
Wiki	0.00 (0.03)	0.00 (0.05)	**0.96** (0.16)	0.00 (0.03)	0.32 (0.44)	**0.75** (0.41)
Wiki-exp	0.00 (0.05)	0.00 (0.07)	**0.78** (0.12)	0.00 (0.02)	0.00 (0.05)	**0.59** (0.40)

  - 일관성 없는 개념과 코로나 관련 개념에 대해 모든 TCAV가 0임 -> 이 개념을 positive와 연관시키지 않음
  - TCAV의 0 점수는 학습 데이터에 해당 개념이 포함되지 않았기 때문일 수도 있다. (Founta의 COVID-19 관련)
  - 1에 가까운 TCAV 점수는 positive 예측을 위한 개념의 중요성을 나타냄
- 이러한 TCAV score가 일반화 성능을 설명할 수 있는가?

Classifier’s Sensitivity to a Concept

- 특정 개념에 대한 평균 TCAV 점수가 일관성 없는 임의 개념의 평균 TCAV 점수와 크게 다른 경우 분류기가 해당 개념에 민감한 것(sensitive)으로 간주

- 실험 결과
  - 일반 분류기가 명시적 코로나 관련 용설에만 민감하다
  - 분류기가 새로운 도메인의 명시적 용설에 더 잘 일반화 됨을 의미
  - 명시적인 코로나 관련 데이터(CH)로 훈련된 분류기는 암시적 용설 개념에 민감하지 않음
  - 명시적, 암시적 코로나 관련 용설 개념에 민감한 분류기는 EA 분류기 뿐이다.
  - 코로나 관련 데이터셋에 훈련된 분류기는 광범위한 혐오를 다루는 일반적인 혐오 개념에 민감하지 않다.

- 이러한 결과는 도메인 내외적으로 더 나은 일반화 가능성을 위해 훈련 데이터에 암시적인 용설 데이터를 포함하는 것의 중요성을 강조한다.

Degree of Explicitness

- 일반 분류기를 업데이트할 때, 암시적 데이터가 더 많은 정보를 제공하고 데이터 증강을 가능케 함
- 정량적 방법 제공
  - TCAV 방법론을 확장해 문장의 DoE(명시성 정도)를 추정함
  - DoE는 **명시적** 개념에 하나의 **명시적** 데이터를 추가해도 평균 TCAV 점수(1에 가까움)에 영향을 미치지 않는다는 아이디어를 기반으로 함.
  - 하지만 추가 데이터가 암시적이라면 아마 모든 CAV의 방향이 변경되고 수정된 개념에 대한 분류기의 민감도가 감소할 것이다.

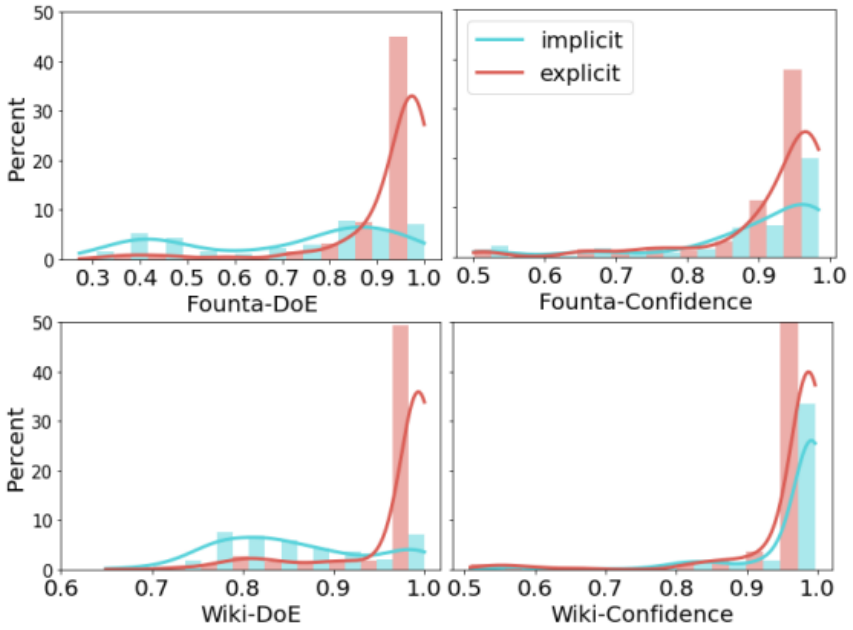
$$v_{new}^{p} = \frac{1}{N_v}\left(\sum_{j=1}^{N_v-1} r_C^j + r_{new}\right), \quad p = 1, .., P$$

  - 데이터 하나를 추가하고 각각의 평균 TCAV 점수를 계산
    - 만약 이 데이터가 명시적이라면 v는 explicit 개념의 표현으로 여겨지고, 평균 TCAV 점수는 1에 가깝게 유지된다.
    - 하지만 이 데이터가 덜 명시적일수록(암시적일수록) 평균 점수가 떨어진다.

DoE analysis on COVID-related abusive data

- 암시적 및 명시적 욕설 데이터를 분류하는 측면에서의 DoE의 유용성 검증 ($N_v == 3$)
- CH 및 EA dev 셋에서 명시적 아시안 차별 개념을 정의하는 데 사용된 예제를 제외하고 암시적 및 명시적 예제의 DoE 점수 계산
- 낮은 분류 신뢰도는 모델이 데이터를 올바르게 예측하는 데 어려움을 겪고 있음을 나타낼 수 있으므로 암시적 데이터는 명시적 데이터보다 분류 신뢰도가 낮을 것으로 예상할 수 있음

- 명시적 데이터와 암시적 데이터의 '분류 신뢰도'는 크게 차이나지 않음
- 그에 반해 'DoE'는 두 그래프가 구분됨
- 즉 DoE가 분류 신뢰도보다 암시적 데이터와 명시적 데이터를 분리하는 데 더 효과적



Data Augmentation with DoE score

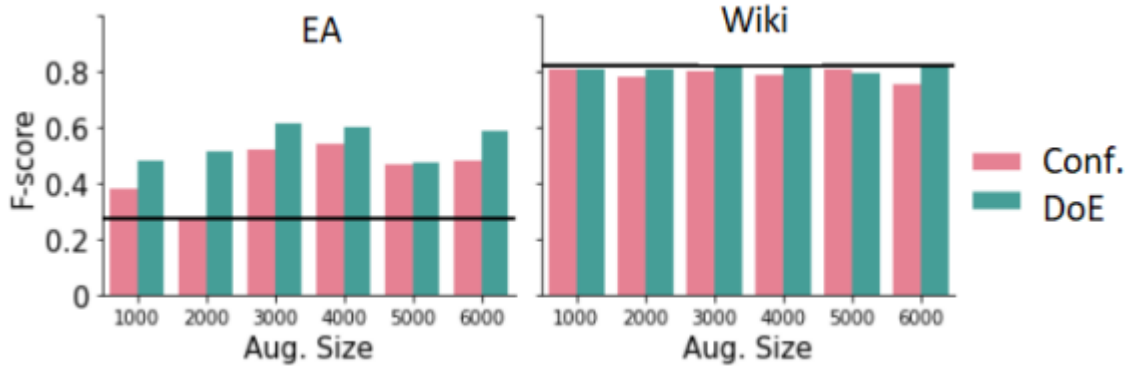
- DoE 점수로 데이터 증강
  - 새로운 도메인의 욕설을 포함하기 위해 일반 분류기가 증강 데이터셋으로 재훈련되어야 하는 것을 고려
  - 앞에서의 실험등을 통해 암시적 데이터가 분류기를 업데이트하는 데 더욱 유익하다는 가설을 고려함
  - DoE 기반 증강 방법을 설명하고 이를 분류 신뢰도 기반 증강 방법과 비교

- 실험
  - 일반적인 Wiki분류기를 사용해 이 분류기가 코로나 관련 아시안 혐오 데이터를 처리할 수 있도록, 원래 wiki 학습셋을 보강하기 위해 **EA 학습셋**의 작지만 충분한 일부를 찾는다.
  - EA 훈련 셋의 모든 데이터에 대한 DoE 및 신뢰도 점수를 계산한다
  - 그 중 wiki 학습 셋에, **가장 낮은 DoE 점수**를 갖는 N개의 예를 더해준다.
    - N은 1k에서 6k까지 다양함 (1k씩)

  - 증강 데이터 크기가 6k가 되면 wiki 테스트 셋의 분류기 성능이 두 기술 모두에서 떨어짐
  - 또한 증강 데이터셋의 크기가 증가함에 따라 두 방법이 동일한 성능으로 수렴됨

Data Augmentation with DoE score

- 아래 그림은 Wiki 테스트셋 및 EA 테스트셋에서 DoE 및 신뢰도 기반 확대 방법을 사용해 업데이트 된 분류기의 f1 점수를 보여줌



  - 검은색 선은 베이스라인을 의미 (기존 Wiki 데이터로 학습 후 각 테스트셋을 평가한 성능)
- EA만이 증강에 사용되기 때문에 이 데이터셋의 분류기를 평가해 증강 학습셋의 최적의 크기를 찾아 가장 성능이 좋은 분류기로 CH를 평가한다.
- 효율적인 증강이 wiki에서는 성능을 유지하고 EA 테스트셋에서는 허용 가능한 결과에 도달해야한다고 예상

Data Augmentation with DoE score

- DoE는 새로운 도메인의 혐오를 학습하는 데 더 좋음
  - EA 데이터셋에서 DoE는 N=5k를 제외하고 모든 N에 대해 신뢰기반 증강 방법보다 더 나은 결과를 얻는다.
- DoE는 기존 데이터셋의 성능을 더 잘 유지함
  - 모든 증강 크기에 대해 DoE 증강 분류기의 성능은 기준선에서 2% 이내로 유지되는 반면, 신뢰 기반 증강의 경우 최대 6%까지 소함
- DoE가 전반적으로 더 좋음
  - 옆의 표는 두가지 증강방법으로 달성한 최상의 결과를 나타냄
  - DoE가 이 데이터셋에 대해서도 더 좋은 성능을 보임
  - 또한 데이터 증강 전후에 분류기의 출력을 질적 평가함
    - 명시적 혐오 발언은 두 경우에 모두 잘 분류됨
    - 하지만 많은 암시적 데이터는 재학습된 분류기에서만 옳게 분류됨
  - "f*ck you china and your chinese virus“ -> 명시적
  - “it is not covid 19 but wuhanvirus” -> 암시적

		F1-score			AUC		
Method	**Aug. set**	***EA***	***CH***	***Wiki***	***EA***	***CH***	***Wiki***
DoE	3K *EA*	**0.61**	**0.73**	**0.82**	**0.81**	**0.78**	**0.96**
Conf.	4K *EA*	0.54	0.71	0.79	0.69	0.75	0.94
Merging	*EA*	0.58	0.72	0.78	0.72	0.75	0.94
baseline	-	0.27	0.69	0.82	0.64	0.74	0.96

Conclusion

- 실제 데이터가 발전함에 따라 학습된 모델이 새로운 데이터에 일반화 됐는지 확인되어야 함
- 민감도를 정량화(수치화)하기 위해 TCAV 도입
  - 이를 사용해 코로나 이전 데이터로 학습된 혐오 분류기의 일반화를 명시적 및 암시적 코로나 관련 아시안 혐오와 비교
- 새로운 도메인에 대한 일반화 가능성을 개선하기 위해 민감도 기반 데이터 증강 접근 방식 제안
  - 가장 유익한 샘플은 '새로운 도메인의 암시적인 데이터' 이다.
- 최적의 개념을 선택하기 위한 전략은 향후 탐색되어야함
- 혐오 탐지뿐만 아니라 감성분석과 같은 다른 NLP에도 적용될 수 있다.
- 언어진화 관점에서 분류기의 행동을 모니터링하고 설명하는 것은 시간이 지날수록 중요해질 것이다.