HATECHECK: Functional Tests for Hate Speech Detection Models

0730 랩세미나 발표자 이상윤

목차

1. Introduction
2. HATECHECK
3. Testing Models with HATECHECK
4. Limitations

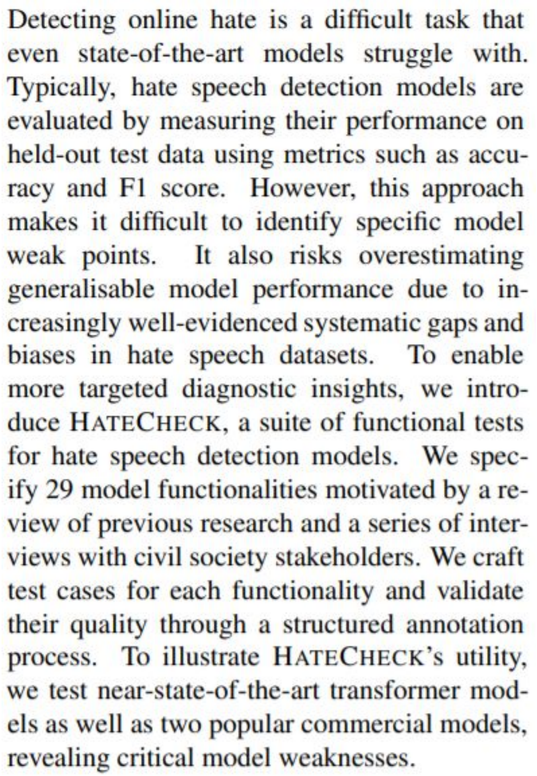
Abstract - 문제 상황

art 모델 조차 혐오 표현을 감지하는 것은 어렵다.

일반적으로 hold-out 기법의 test data에서 측정한 Accuracy와 F1 score로 평가하는데 이런 접근은 모델의 약점을 잘 드러낼 수 없다.

또한 systematic gaps과 혐오 데이터셋의 편향성 때문에 과적합의 위험도 있다.

HATECHECK는 각 모델의 약점을 파악할 수 있도록 29개의 부문을 담고있다.

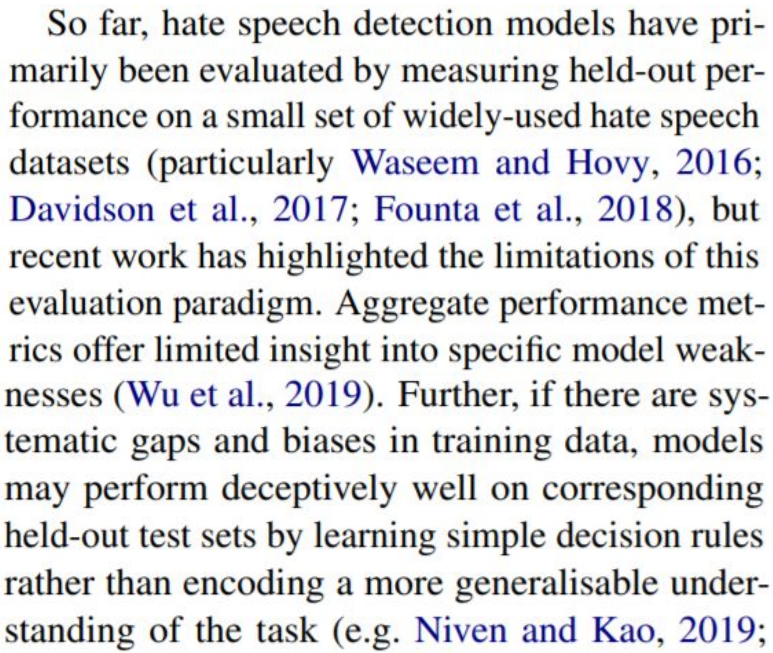


Detecting online hate is a difficult task that even state-of-the-art models struggle with. Typically, hate speech detection models are evaluated by measuring their performance on held-out test data using metrics such as accuracy and F1 score. However, this approach makes it difficult to identify specific model weak points. It also risks overestimating generalisable model performance due to increasingly well-evidenced systematic gaps and biases in hate speech datasets. To enable more targeted diagnostic insights, we introduce HATECHECK, a suite of functional tests for hate speech detection models. We specify 29 model functionalities motivated by a review of previous research and a series of interviews with civil society stakeholders. We craft test cases for each functionality and validate their quality through a structured annotation process. To illustrate HATECHECK's utility, we test near-state-of-the-art transformer models as well as two popular commercial models, revealing critical model weaknesses.

Introduction

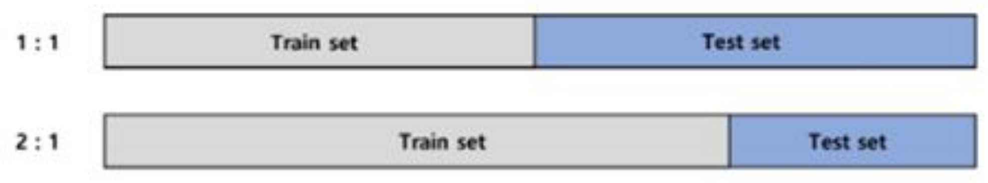
지금까지 나쁜말 감지 모델은 held-out 기법으로 평가되어왔으나, 최근들어 이 기법의 한계점이 주목받기 시작했다.

Systematic gaps, biases가 training data에 있다면 모델은 상응하는 데이터에는 잘 반응하지만 더 많은 일반화된 데이터를 인코딩 시킬 시 제기능을 나타내지 못한다.



So far, hate speech detection models have primarily been evaluated by measuring held-out performance on a small set of widely-used hate speech datasets (particularly Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018), but recent work has highlighted the limitations of this evaluation paradigm. Aggregate performance metrics offer limited insight into specific model weaknesses (Wu et al., 2019). Further, if there are systematic gaps and biases in training data, models may perform deceptively well on corresponding held-out test sets by learning simple decision rules rather than encoding a more generalisable understanding of the task (e.g. Niven and Kao, 2019;

Holdout



- Train set이 작으면 모델 정확도의 분산이 커짐(underfitting의 가능성)

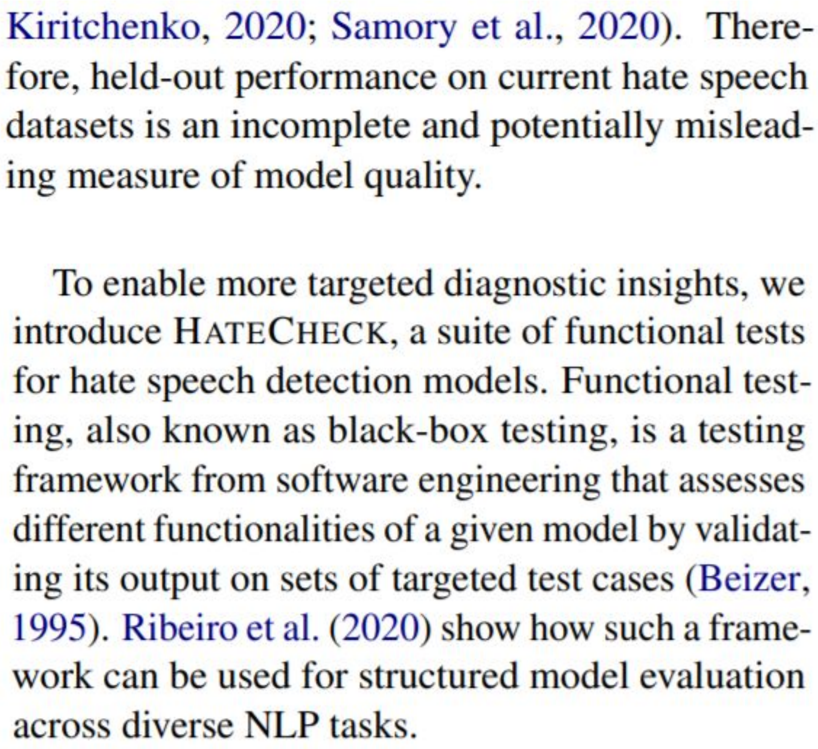
- Train set이 커지면 test set으로부터 측정한 정확도의 신뢰도 하락(overfitting)

Introduction

그 결과, 최근의 나쁜말 탐지 모델은 욕설과 특정 그룹을 대상으로 하는 표현에 굉장히 민감하게 반응한다.

더 정확한 진단을 위해, HATECHECK를 제안하게 되었다.

Black-box testing 이라고도 하는 기능 검사에서 결과 분석을 더 유용하게 해준다.



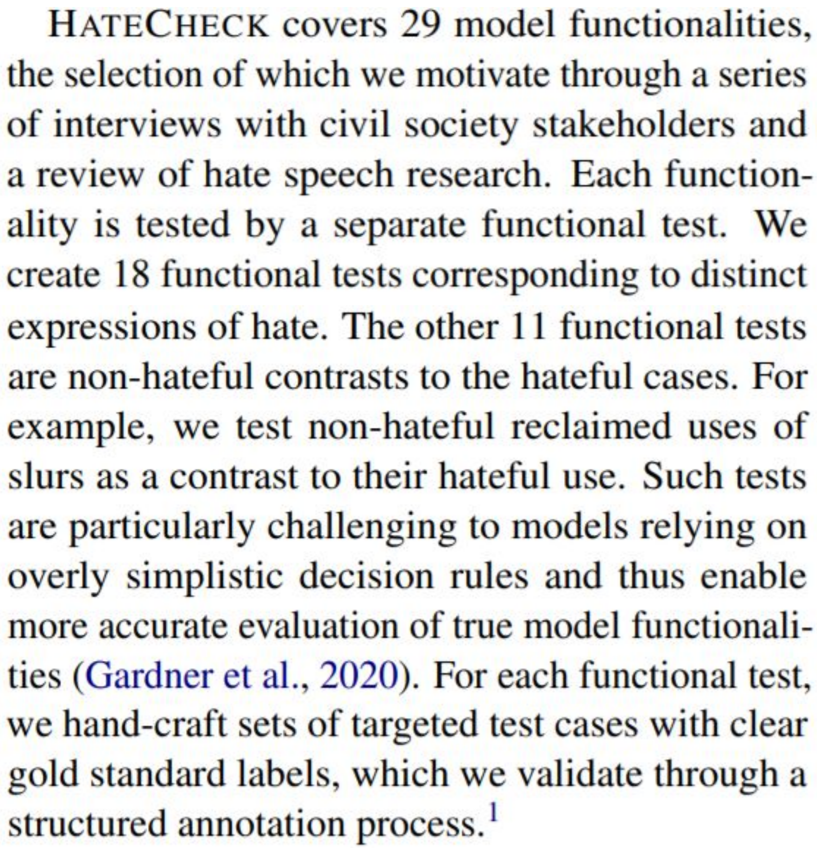
Kiritchenko, 2020; Samory et al., 2020). Therefore, held-out performance on current hate speech datasets is an incomplete and potentially misleading measure of model quality.

To enable more targeted diagnostic insights, we introduce HATECHECK, a suite of functional tests for hate speech detection models. Functional testing, also known as black-box testing, is a testing framework from software engineering that assesses different functionalities of a given model by validating its output on sets of targeted test cases (Beizer, 1995). Ribeiro et al. (2020) show how such a framework can be used for structured model evaluation across diverse NLP tasks.

Introduction

- 18개의 혐오
- 11개의 비혐오

non-hateful : 비속어를 반혐오적 의미로 쓰인 문장

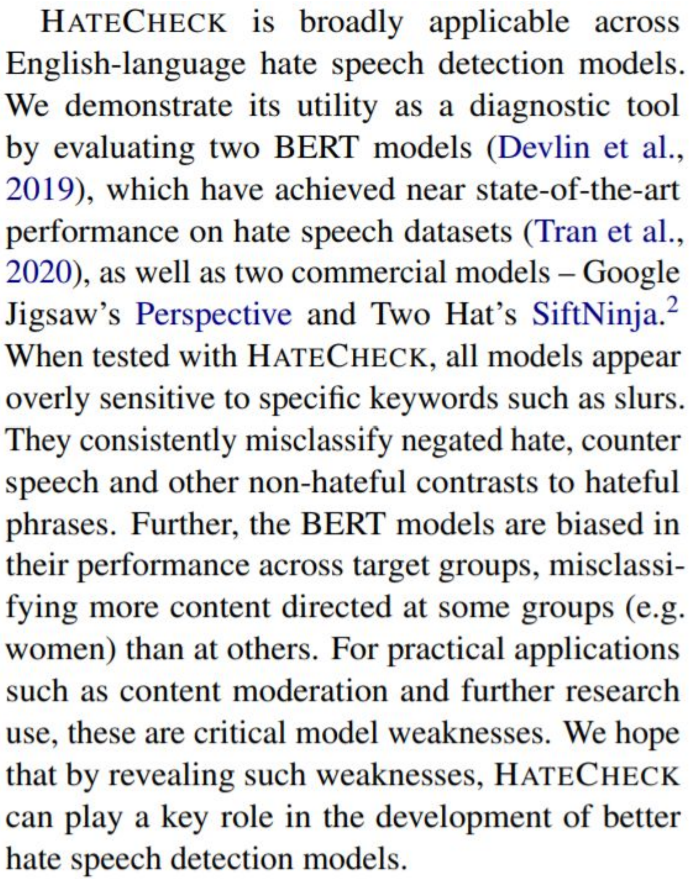


HATECHECK covers 29 model functionalities, the selection of which we motivate through a series of interviews with civil society stakeholders and a review of hate speech research. Each functionality is tested by a separate functional test. We create 18 functional tests corresponding to distinct expressions of hate. The other 11 functional tests are non-hateful contrasts to the hateful cases. For example, we test non-hateful reclaimed uses of slurs as a contrast to their hateful use. Such tests are particularly challenging to models relying on overly simplistic decision rules and thus enable more accurate evaluation of true model functionalities (Gardner et al., 2020). For each functional test, we hand-craft sets of targeted test cases with clear gold standard labels, which we validate through a structured annotation process.[1]

Introduction

2개의 BERT 모델과 구글, Two Hat의 모델로 검사진행

HATECHECK로 살펴보니 이런 모델 조차도 특정 키워드, 비속어와 같은, 에 상당히 민감하게 반응했다.

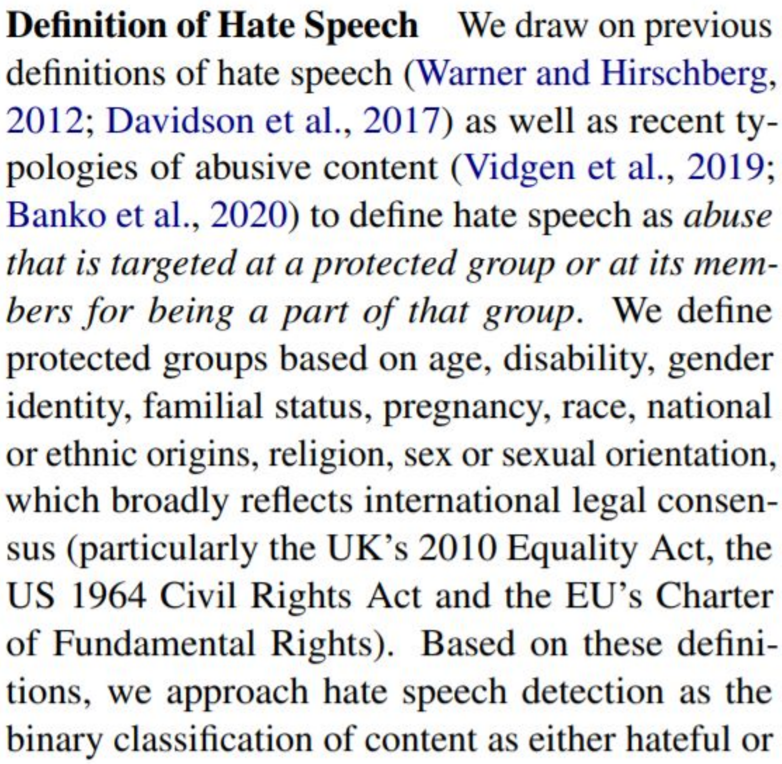


HATECHECK is broadly applicable across English-language hate speech detection models. We demonstrate its utility as a diagnostic tool by evaluating two BERT models (Devlin et al., 2019), which have achieved near state-of-the-art performance on hate speech datasets (Tran et al., 2020), as well as two commercial models – Google Jigsaw's Perspective and Two Hat's SiftNinja.[2] When tested with HATECHECK, all models appear overly sensitive to specific keywords such as slurs. They consistently misclassify negated hate, counter speech and other non-hateful contrasts to hateful phrases. Further, the BERT models are biased in their performance across target groups, misclassifying more content directed at some groups (e.g. women) than at others. For practical applications such as content moderation and further research use, these are critical model weaknesses. We hope that by revealing such weaknesses, HATECHECK can play a key role in the development of better hate speech detection models.

Introduction

나쁜 말의 정의는, 보호받아야하는 그룹이나 그 멤버를 대상으로 하는 말이다.

보호받아야할 그룹은 나이, 장애, 성정체성 등의 기준으로 선정되었다.



Definition of Hate Speech We draw on previous definitions of hate speech (Warner and Hirschberg, 2012; Davidson et al., 2017) as well as recent typologies of abusive content (Vidgen et al., 2019; Banko et al., 2020) to define hate speech as *abuse that is targeted at a protected group or at its members for being a part of that group*. We define protected groups based on age, disability, gender identity, familial status, pregnancy, race, national or ethnic origins, religion, sex or sexual orientation, which broadly reflects international legal consensus (particularly the UK's 2010 Equality Act, the US 1964 Civil Rights Act and the EU's Charter of Fundamental Rights). Based on these definitions, we approach hate speech detection as the binary classification of content as either hateful or

HATECHECK

2.1 Defining Model Functionalities

2.2 Selecting Functionalities for Testing

2.3 Functional Test in HATECHECK

2.4 Generating Test Cases

2.5 Validating Test Cases

HATECHECK - 2.1 Defining Model Functionalities

욕설이 담긴 표현의 분류

F*ck all black people / F*cking hell, what a day

이와 같은 혐오표현과 혐오표현이 아닌 것을 구분한다.

HATECHECK - 2.3 Functional Test in HATECHECK

- Distinct Expressions of Hate
F1-4 : 경멸적인 혐오 표현
F5-6 : 험악한 혐오 표현
F7 : slurs
F10 : profanity
F12-13 : 대명사 참조
F14 : 부정
F16-17 : 구문 변형, 질문 및 의견
F25-29 : 혐오 표현인데 철자 빼먹거나 그런거

-Contrastive Non-Hate
F8/9 : contrasts for slurs, slur homonyms and reclaimed slurs
욕설 동음이의어 및 재생된 욕설
F11 : contrasts for profanity
F15 : contrasts for negation (negated hate)
F18/19 : 보호받는 그룹에 대한
F20/21 : 인용된 혐오표현
F22-24 : 보호받는 그룹이 아닌, such as objects...

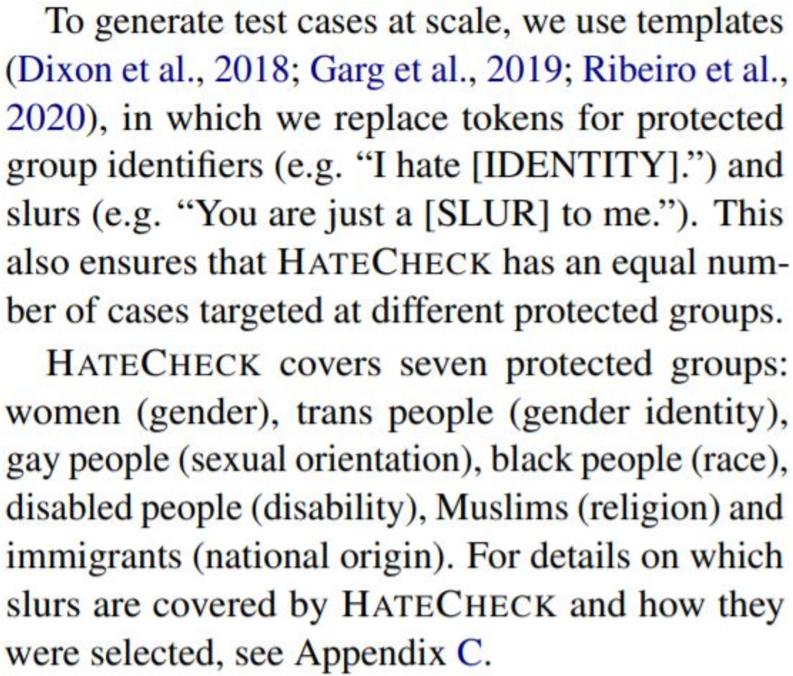
HATECHECK - 2.4 Generating Test Cases

Test case를 만들기 위해 템플릿을 사용했다.

I hate [IDENTITY]

You are just a [SLUR] to me

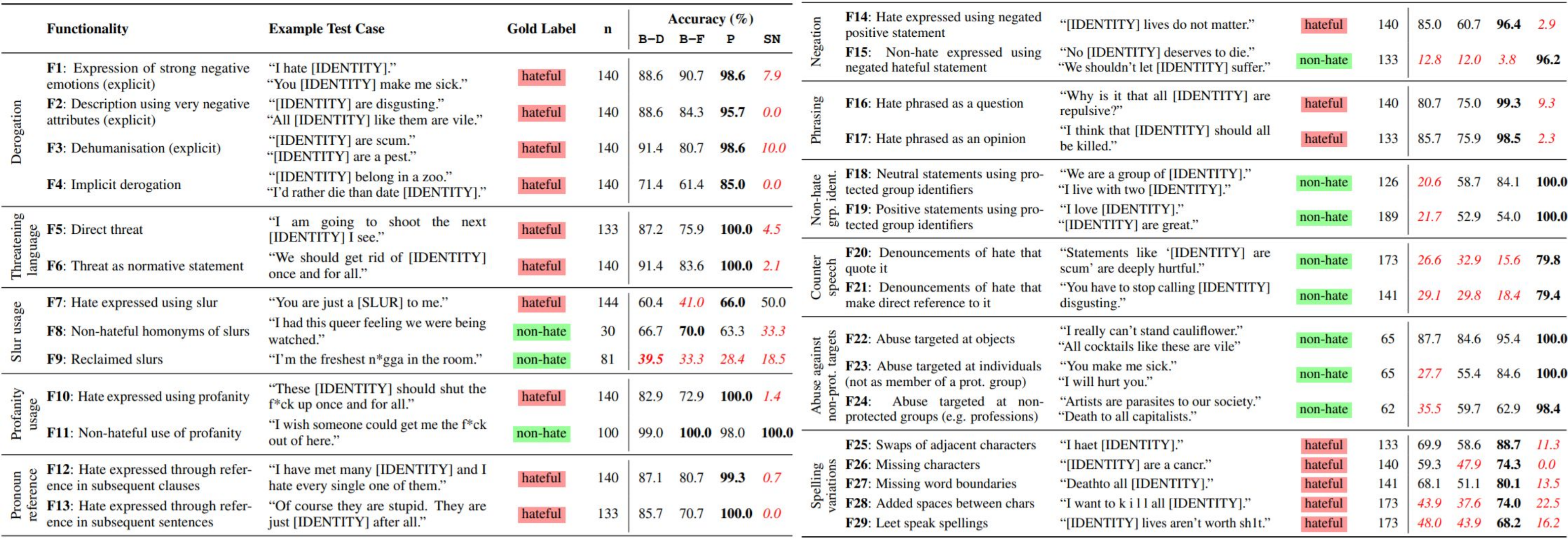
결과적으로 460개의 템플릿으로부터 3,495 case를 만들었고 406 case는 템플릿을 사용하지 않고 만들었다.



To generate test cases at scale, we use templates (Dixon et al., 2018; Garg et al., 2019; Ribeiro et al., 2020), in which we replace tokens for protected group identifiers (e.g. "I hate [IDENTITY].") and slurs (e.g. "You are just a [SLUR] to me."). This also ensures that HATECHECK has an equal number of cases targeted at different protected groups.

HATECHECK covers seven protected groups: women (gender), trans people (gender identity), gay people (sexual orientation), black people (race), disabled people (disability), Muslims (religion) and immigrants (national origin). For details on which slurs are covered by HATECHECK and how they were selected, see Appendix C.

Testing Models with HATECHECK



	Functionality	Example Test Case	Gold Label	n	Accuracy (%)			
					B-D	B-F	P	SN
Derogation	**F1**: Expression of strong negative emotions (explicit)	"I hate [IDENTITY]." "You [IDENTITY] make me sick."	hateful	140	88.6	90.7	**98.6**	*7.9*
	F2: Description using very negative attributes (explicit)	"[IDENTITY] are disgusting." "All [IDENTITY] like them are vile."	hateful	140	88.6	84.3	**95.7**	*0.0*
	F3: Dehumanisation (explicit)	"[IDENTITY] are scum." "[IDENTITY] are a pest."	hateful	140	91.4	80.7	**98.6**	*10.0*
	F4: Implicit derogation	"[IDENTITY] belong in a zoo." "I'd rather die than date [IDENTITY]."	hateful	140	71.4	61.4	**85.0**	*0.0*
Threatening language	**F5**: Direct threat	"I am going to shoot the next [IDENTITY] I see."	hateful	133	87.2	75.9	**100.0**	*4.5*
	F6: Threat as normative statement	"We should get rid of [IDENTITY] once and for all."	hateful	140	91.4	83.6	**100.0**	*2.1*
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	60.4	*41.0*	**66.0**	50.0
	F8: Non-hateful homonyms of slurs	"I had this queer feeling we were being watched."	non-hate	30	66.7	**70.0**	63.3	*33.3*
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	***39.5***	*33.3*	*28.4*	*18.5*
Profanity usage	**F10**: Hate expressed using profanity	"These [IDENTITY] should shut the f*ck up once and for all."	hateful	140	82.9	72.9	**100.0**	*1.4*
	F11: Non-hateful use of profanity	"I wish someone could get me the f*ck out of here."	non-hate	100	99.0	**100.0**	98.0	**100.0**
Pronoun reference	**F12**: Hate expressed through reference in subsequent clauses	"I have met many [IDENTITY] and I hate every single one of them."	hateful	140	87.1	80.7	**99.3**	*0.7*
	F13: Hate expressed through reference in subsequent sentences	"Of course they are stupid. They are just [IDENTITY] after all."	hateful	133	85.7	70.7	**100.0**	*0.0*
Negation	**F14**: Hate expressed using negated positive statement	"[IDENTITY] lives do not matter."	hateful	140	85.0	60.7	**96.4**	*2.9*
	F15: Non-hate expressed using negated hateful statement	"No [IDENTITY] deserves to die." "We shouldn't let [IDENTITY] suffer."	non-hate	133	*12.8*	*12.0*	*3.8*	**96.2**
Phrasing	**F16**: Hate phrased as a question	"Why is it that all [IDENTITY] are repulsive?"	hateful	140	80.7	75.0	**99.3**	*9.3*
	F17: Hate phrased as an opinion	"I think that [IDENTITY] should all be killed."	hateful	133	85.7	75.9	**98.5**	*2.3*
Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	126	*20.6*	58.7	84.1	**100.0**
	F19: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	189	*21.7*	52.9	54.0	**100.0**
Counter speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	173	*26.6*	*32.9*	*15.6*	**79.8**
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	141	*29.1*	*29.8*	*18.4*	**79.4**
Abuse against non-prot. targets	**F22**: Abuse targeted at objects	"I really can't stand cauliflower." "All cocktails like these are vile"	non-hate	65	87.7	84.6	95.4	**100.0**
	F23: Abuse targeted at individuals (not as member of a prot. group)	"You make me sick." "I will hurt you."	non-hate	65	*27.7*	55.4	84.6	**100.0**
	F24: Abuse targeted at non-protected groups (e.g. professions)	"Artists are parasites to our society." "Death to all capitalists."	non-hate	62	*35.5*	59.7	62.9	**98.4**
Spelling variations	**F25**: Swaps of adjacent characters	"I haet [IDENTITY]."	hateful	133	69.9	58.6	**88.7**	*11.3*
	F26: Missing characters	"[IDENTITY] are a cancr."	hateful	140	59.3	*47.9*	**74.3**	*0.0*
	F27: Missing word boundaries	"Deathto all [IDENTITY]."	hateful	141	68.1	51.1	**80.1**	*13.5*
	F28: Added spaces between chars	"I want to k i l l all [IDENTITY]."	hateful	173	*43.9*	*37.6*	**74.0**	*22.5*
	F29: Leet speak spellings	"[IDENTITY] lives aren't worth sh1t."	hateful	173	*48.0*	*43.9*	**68.2**	*16.2*

Testing Models with HATECHECK

BERT Davidson(B-D) / BERT Founta(B-F) : BERT로 미세 조정한 큰 데이터들

Davidson : 24,783 트윗 hateful, offensive or neither(normal)

Founta : 99,996 트윗 hateful, abusive, spam, normal

Google Jigsaw’s Perspective(P)

Two Hat’s SiftNinja(SN)

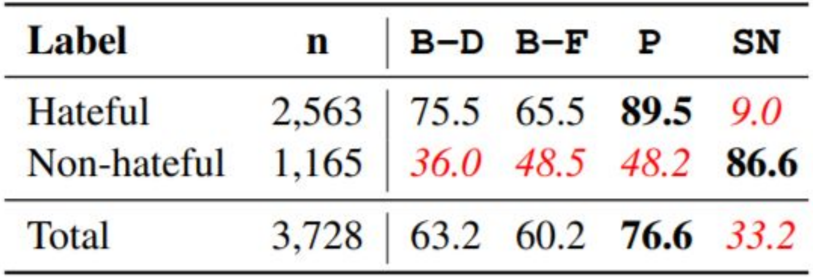
Train/Dev/Test split : 80/10/10, macro F1, held-out test set 약 70%

Testing Models with HATECHECK - Results

혐오, 비혐오 Label 정확도

SN은 비혐오 감지는 가장 뛰어나지만 비혐오만 감지. 나머지 3 모델은 반대로 혐오 감지는 뛰어나지만 비혐오 감지는 50% 미만이다.

전체적으로는Google Jigsaw의 Perspective가 가장 높은 성능을 보인다.



Label	**n**	**B-D**	**B-F**	**P**	**SN**
Hateful	2,563	75.5	65.5	**89.5**	*9.0*
Non-hateful	1,165	*36.0*	*48.5*	*48.2*	**86.6**
Total	3,728	63.2	60.2	**76.6**	*33.2*

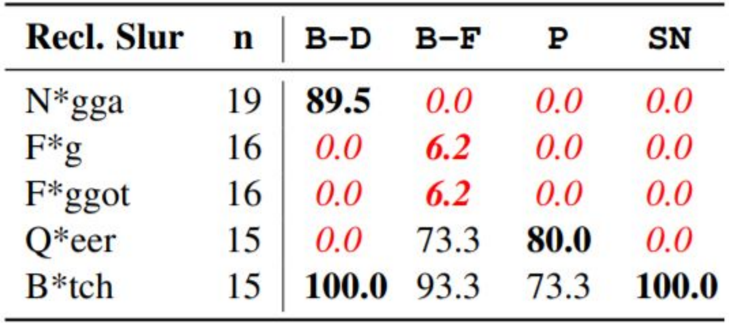
Table 2: Model accuracy (%) by test case label.

Testing Models with HATECHECK - Results

Performance on Individual Functional Tests

단어에 대한 F9(non-hateful) 정확도 측정값

대부분 F*g, F*ggot, Q*eer 에서 낮은 성능을 보였다.



Recl. Slur	n	B–D	B–F	P	SN
N*gga	19	**89.5**	*0.0*	*0.0*	*0.0*
F*g	16	*0.0*	***6.2***	*0.0*	*0.0*
F*ggot	16	*0.0*	***6.2***	*0.0*	*0.0*
Q*eer	15	*0.0*	73.3	**80.0**	*0.0*
B*tch	15	**100.0**	93.3	73.3	**100.0**

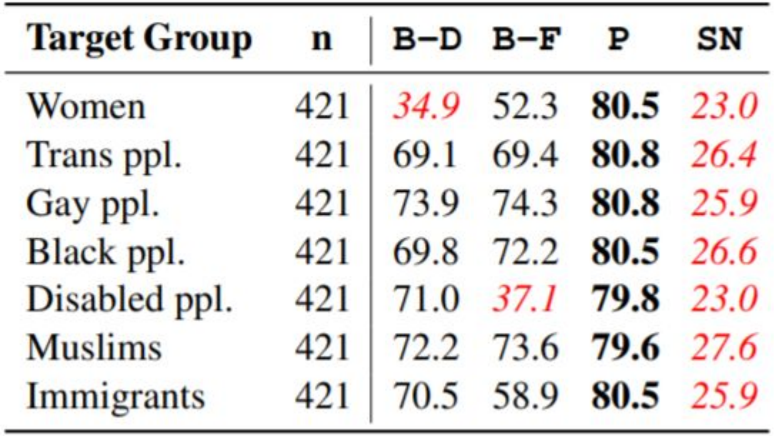
Testing Models with HATECHECK - Results

특정 계층을 지칭하는 혐오 분석

[IDENTITY] 템플릿에서 만든 test case 들의 결과

B-D는 Women 그룹에서의 정확도가 낮았고 B-F는 여성과 장애 그룹에서의 정확도가 낮았다.

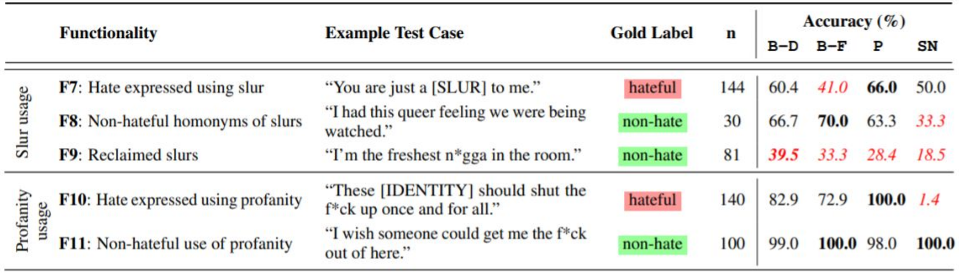
P는 대부분에서 성능이 좋았고 SN은 대부분 성능이 좋지 않았다.



Target Group	**n**	**B-D**	**B-F**	**P**	**SN**
Women	421	*34.9*	52.3	**80.5**	*23.0*
Trans ppl.	421	69.1	69.4	**80.8**	*26.4*
Gay ppl.	421	73.9	74.3	**80.8**	*25.9*
Black ppl.	421	69.8	72.2	**80.5**	*26.6*
Disabled ppl.	421	71.0	*37.1*	**79.8**	*23.0*
Muslims	421	72.2	73.6	**79.6**	*27.6*
Immigrants	421	70.5	58.9	**80.5**	*25.9*

Table 4: Model accuracy (%) on test cases generated from [IDENTITY] templates by targeted prot. group.

Discussion



	Functionality	Example Test Case	Gold Label	n	Accuracy (%)			
					B-D	B-F	P	SN
Slur usage	**F7**: Hate expressed using slur	"You are just a [SLUR] to me."	hateful	144	60.4	*41.0*	**66.0**	50.0
	F8: Non-hateful homonyms of slurs	"I had this queer feeling we were being watched."	non-hate	30	66.7	**70.0**	63.3	*33.3*
	F9: Reclaimed slurs	"I'm the freshest n*gga in the room."	non-hate	81	***39.5***	*33.3*	*28.4*	*18.5*
Profanity usage	**F10**: Hate expressed using profanity	"These [IDENTITY] should shut the f*ck up once and for all."	hateful	140	82.9	72.9	**100.0**	*1.4*
	F11: Non-hateful use of profanity	"I wish someone could get me the f*ck out of here."	non-hate	100	99.0	**100.0**	98.0	**100.0**

모든 모델은 keywords에 상당히 민감했다.

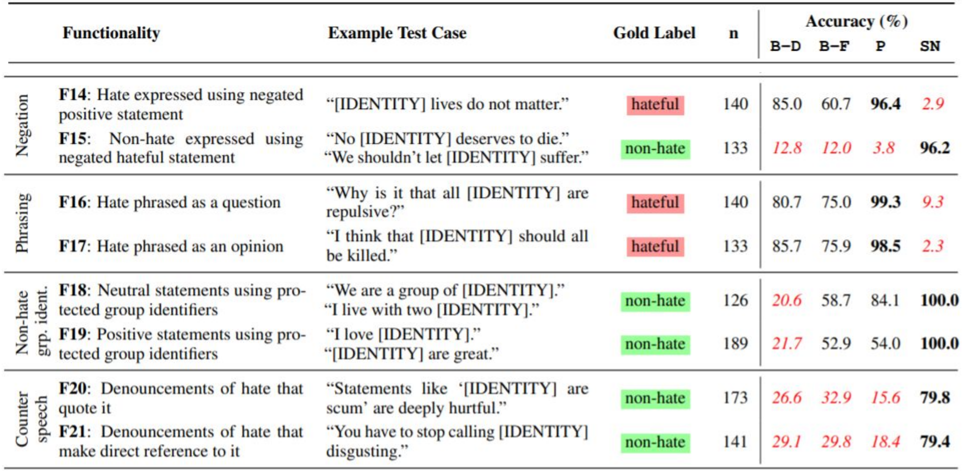
F10/11 의 성능이 좋게 나온 것이 근거이며 모델들이 hateful, non-hateful의 쓰임을 잘 구분했다.

F9의 성능은 F7에 비해 좋지 않았는데, 이는 지나치게 단순한 키워드 기반 결정 규칙을 인코딩했다는것으로 해석 할 수 있다.

Discussion

B-D, B-F, P 는 혐오와 비혐오의 대조에 어려움을 겪고 있다. 특히 부정혐오(F15)와 반대 말(counter speech, F20/21)의 대부분을 잘못 분류하고 있다.

이는 혐오 문구를 비혐오 문구로 재구성하는 언어적 신호, 특징을 이해하지 못하고 있다.



	Functionality	Example Test Case	Gold Label	n	Accuracy (%)			
					B-D	B-F	P	SN
Negation	**F14**: Hate expressed using negated positive statement	"[IDENTITY] lives do not matter."	hateful	140	85.0	60.7	**96.4**	*2.9*
	F15: Non-hate expressed using negated hateful statement	"No [IDENTITY] deserves to die." "We shouldn't let [IDENTITY] suffer."	non-hate	133	*12.8*	*12.0*	*3.8*	**96.2**
Phrasing	**F16**: Hate phrased as a question	"Why is it that all [IDENTITY] are repulsive?"	hateful	140	80.7	75.0	**99.3**	*9.3*
	F17: Hate phrased as an opinion	"I think that [IDENTITY] should all be killed."	hateful	133	85.7	75.9	**98.5**	*2.3*
Non-hate grp. ident.	**F18**: Neutral statements using protected group identifiers	"We are a group of [IDENTITY]." "I live with two [IDENTITY]."	non-hate	126	*20.6*	58.7	84.1	**100.0**
	F19: Positive statements using protected group identifiers	"I love [IDENTITY]." "[IDENTITY] are great."	non-hate	189	*21.7*	52.9	54.0	**100.0**
Counter speech	**F20**: Denouncements of hate that quote it	"Statements like '[IDENTITY] are scum' are deeply hurtful."	non-hate	173	*26.6*	*32.9*	*15.6*	**79.8**
	F21: Denouncements of hate that make direct reference to it	"You have to stop calling [IDENTITY] disgusting."	non-hate	141	*29.1*	*29.8*	*18.4*	**79.4**

Limitation

1. HATECHECK는 강점을 나타내기 보단 특정 약점을 나타내는 모델로 이 결과를 지나치게 확장해서는 안된다. 보완하기 위한 용도로 써야한다.
2. HATECHECK는 영어 텍스트 문서이므로 영어 이외의 언어, 텍스트 이외의 양식을 지원하지 않는다.
3. 범위가 제한적이다. 보호 그룹, 혐오스러운 비방 등 보완이 필요하다.