

Ruddit: Norms of Offensiveness for English Reddit Comments

윤예준

INDEX

01 연구 배경

02 Related Work

03 Data collection and sampling

04 Annotation

05 Data Analysis

06 Computational Modeling

07 Conclusion

01

연구 배경



그림 1. 소셜 미디어
플랫폼

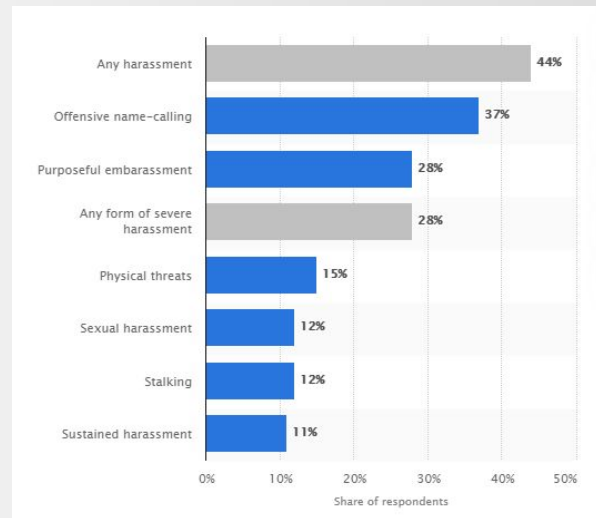


그림 2. 2020년 1월 기준으로
개인적으로
온라인 괴롭힘을 경험한 미국 성인
인터넷 사용자 비율

Amily Munro. 2011. The protection of children online:
A brief scoping review to identify vulnerable groups

Waseem and Hovy (2016)

– Racist, sexist

Davidson et al. (2017)

– Hate-speech, offensive, offensive but not
hate-speech

Founta et al. (2018)

– Abusive, hateful, normal, spam

Dataset	# Tweets	Labels	Annotators
(Chatzakou et al. 2017)	9,484	aggressive, bullying, spam, normal	5
(Waseem and Hovy 2016)	16,914	racist, sexist, normal	1
(Davidson et al. 2017)	24,802	hateful, offensive (but not hateful), neither	3 or more
(Golbeck et al. 2017)	35,000	the worst, threats, hate speech, direct harassment, potentially offensive, non-harassment	2-3
Present study	80,000	offensive, abusive, hateful speech, aggressive, cyberbullying, spam, normal	5-20

그림 3. Founta et al. (2018)

1 ABUSIVE
2 alay
3 ampas
4 buta
5 keparat
6 anjing
7 anjir
8 babi
9 bacot
10 bajingan
11 banci
12 bandot
13 buaya
14 bangkai
15 bangsat
16 bego

그림 4. Davidson et al. (2017)

Surface Realisation Using Full Delexicalisation

그림 5. Gao and Huang, 2017

02


Related Work

- Offensive Language Datasets
- Best–Worst Scaling (BWS)

03

Data collection and sampling

- Topics (50%): AskMen, AskReddit, TwoXChromosomes, vaxxhappened, worldnews, worldpolitics
- ChangeMyView (25%): controversial topics
- Random (25%): random subreddits

- Increase the proportion of offensive and emotional comments. 

- Valence, arousal, dominance (VAD)

The NRC VAD Lexicon

term	$\frac{A}{Z}$	valence	arousal	dominance
aaaaaaah		0.479	0.606	0.291
aaaah		0.520	0.636	0.282
aardvark		0.427	0.490	0.437
aback		0.385	0.407	0.288
abacus		0.510	0.276	0.485
abalone		0.500	0.480	0.412

⌘ 1. NRC VAD Lexicon

04

Annotation

Capturing Reliable Fine-Grained Sentiment Associations by Crowdsourcing and Best–Worst Scaling

Scripts (last updated May, 2017): Code to assist with best-worst-scaling annotations can be downloaded by clicking [here](#). It includes a script to produce 4-tuples with desired term distributions, a script to produce real-valued scores from best-worst annotations, as well as a script to calculate [split-half reliability of the annotations](#).

그림 6. Best–Worst Scaling
(saifmohammad.com)

Focus terms:

1. th*nks 2. doesn't work 3. w00t 4. #theworst

Q1: Identify the term that is associated with the most amount of positive sentiment (or least amount of negative sentiment) – **the most positive term**:

1. th*nks 2. doesn't work 3. w00t 4. #theworst

Q2: Identify the term that is associated with the most amount of negative sentiment (or least amount of positive sentiment) – **the most negative term**:

1. th*nks 2. doesn't work 3. w00t 4. #theworst

그림 7. 위 논문의 4-tuple 질문 예시

Social media conversations often include an original post followed by other comments in response to the original post. These response comments can vary from being highly supportive and pleasant to highly offensive and abusive. We are interested in determining the degree of offensiveness/supportiveness of response comments. Specifically, in each HIT, you will be given four comments. Your task is to mark:

- the comment that is the **MOST offensive** (LEAST supportive).
- the comment that is the **LEAST offensive** (MOST supportive).

Purpose of the task:

Your responses will be used to better understand the type, nature and degree of offensive and supportive language.

Notes:

- The comments included in this task may contain foul, abusive language (including but not limited to sexist, violent, vulgar or racist terms).
- Try not to overthink the answer. *Let your instinct guide you.*
- There are a certain number of gold annotations interspersed throughout the data. If you mark these gold questions incorrectly, we will give you immediate feedback in a pop-up box. We will keep track of your answers for these gold questions. *If you mark too many of these incorrectly, it might lead to rejection of your HITs, upon further scrutiny.*

Instructions:

- For the MOST offensive (LEAST supportive) question: If all four comments are not offensive, then select the comment that is most negative or least supportive. If **two (or more) comments are equally offensive**, and both are MOST offensive (among the four options), then choose any one of these equally offensive comments as the MOST offensive comment.
- For the LEAST offensive (MOST supportive) question: If all four comments are not supportive, then select the comment that is most positive or least offensive. If **two (or more) comments are equally supportive**, and both are MOST supportive (among the four options), then choose any one of these equally supportive comments as the MOST supportive comment.
- Consider 'offensive comments' to be a broad category that includes abhorrent, abusive, rude, objectionable, disagreeable, and discourteous comments. Offensive comments often denigrate, criticize, hurt, or harm the perception of self of another person or group (with or without using offensive words). Offensive comments tend to include aggressive outbursts, name calling, anger, hostility, bullying, or hurtful sarcasm, but they can be more subtle as well.
- Consider 'supportive comments' to be a broad category that includes agreeable, delightful, kind, nice, pleasant, pleasing, respectful, and soothing comments.
- You cannot pick the same comment for both the categories.
- Responses that are not in accordance with the instructions will not be paid for.

그림 8. 주석 지침

Q. From the four comments below, choose the comment which is **LEAST OFFENSIVE** (most supportive) and the comment which is **MOST OFFENSIVE** (least supportive).

Least Offensive	Comment	Most Offensive
<input type="radio"/>	The first time I had sex the guy took off the condom and didn't tell me. Men do this shit too. I don't care about how you feel about it, shitty people exist in both genders.	<input type="radio"/>
<input type="radio"/>	Promo girls is just a euphemism for escort right Cos I m seeing a theme tonight	<input type="radio"/>
<input type="radio"/>	I've heard of quite a few guys having sex with girls who were lying about being on birth control, and had to live with it for the rest of their lives. This does happen quite often	<input type="radio"/>
<input type="radio"/>	I'm clearly not as good at this as I think I am. That being said it seems like there's a very intentional effort to conflate (a) "I was just sitting in a known gang area and a fight broke out and I got hit" with (b) a woman was dragged behind a dumpster and raped because (a) is more common than (b) which obscures the fact that (b) happens almost exclusively to women.	<input type="radio"/>

Figure 7: Sample questionnaire for the final annotation task.

그림 9. 샘플 설문지

Split-half reliability (SHR)

# Comments	# Annotations per Tuple	# Annotations	# Annotators	SHR Pearson	SHR Spearman
6000	6	95,255	725	0.8818 ± 0.0023	0.8612 ± 0.0029

Table 1: Ruddit annotation statistics and split-half reliability (SHR) scores.

표 2. 주석에 대한 SHR

05

Data Analysis

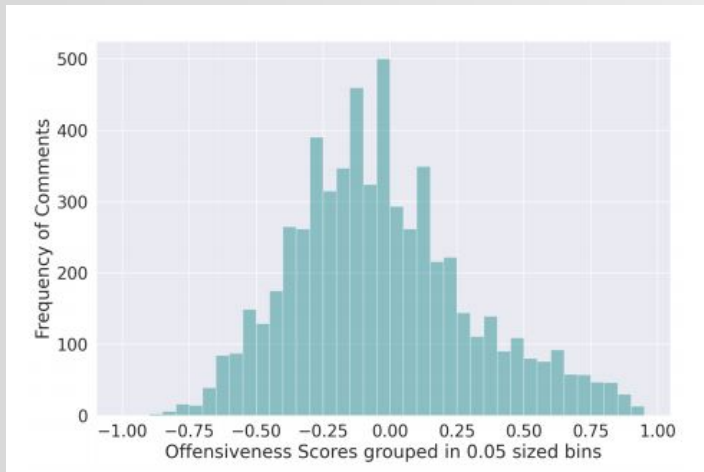


그림 10. 댓글 빈도 - 공격성 정도 (bin 0.05)

Bin	Comment	Score
1	Don't worry, she's going to be fine.	-0.75
	I see you too are a man of culture;)	-0.604
2	This is so sexy! Love it!	-0.562
	"I live with my ex, but it's totally cool, we're just friends"	-0.229
3	Not sure why Im being down voted? Why does the truth bother so many people?	-0.191
	I presented a hypothetical question to you. I did not even claim that you made that argument. Unfortunately that is not a straw man. So, care to answer that question again?	0.083
4	Don't forget Vaccines cause autism. And torture is awesome. We should murder the families of terrorists.	0.5
	What is your angle, Kim?? Is this some Hitler BS where you sign a peace treaty and then start WWII? Or did you finally just grow a brain? Because neither sound particularly more likely than the other...	0.521
5	If you support trump kill yourself, painfully	0.604
	shut the fuck up bitch. It's Bernie or Bust nobody is voting for Biden, now get the fuck out of here you cunt	0.958

표 3. 데이터 세트의 일부 주석

데이터 분석을 위해 0.4 크기의 bin 5개를 주석에 배치 (bin 1: -1.0 ~ -0.6, bin 2: -0.6 ~ -0.2 ...)

PMI(Pointwise Mutual Information)

Bin	Words
1	awesome, thanks, appreciate
2	songs, headphones, sweet, movie
3	gap, sacrifice, employee
4	muslim, fucked, gay, ass, raped
5	dick, fuck, asshole, ass, shut

표 4. 상위 PMI 점수 단어

The terms used are *lesbian, gay, bisexual, transgender, trans, queer, lgbt, lgbtq, homosexual, straight, heterosexual, male, female, nonbinary, african, africanamerican, black, white, european, hispanic, latino, latina, latinx, mexican, canadian, american, asian, indian, middle eastern, chinese, japanese, christian, muslim, jewish, buddhist, catholic, protestant, sikh, taoist, old, older, young, younger, teenage, millennial, middle aged, elderly, blind, deaf, paralyzed, atheist, feminist, islam, muslim, man, woman, boy, girl.*

그림 11. 부록 A.4에 나오는 정체성 용어

identity-agnostic dataset

Emotion	Pearson's r
High Valence	0.0365
Low Valence	0.2140
High Arousal	0.3562
Low Arousal	0.0859
High Dominance	0.0755
Low Dominance	0.1004

Table 4: Pearson correlation values between the offensiveness scores and the emotion dimension scores.

표 4. 감정과 공격성의 상관관계

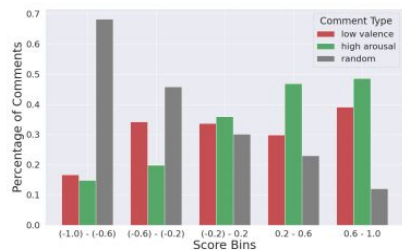


Figure 3: Distribution of comments within each of the 5 score bins over the comment types.

그림 13. bin 내부의 댓글 유형 분포

높은 강도의 V/A/D 단어: valence, arousal, dominance > 0.75

낮은 강도의 V/A/D 단어: valence, arousal, dominance < 0.25

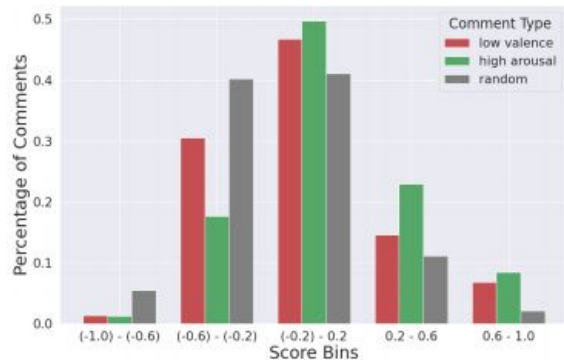


Figure 2: Distribution of comments in each comment type over the 5 offensiveness score bins.

그림 12. 댓글 유형에 따른 댓글 분포

Dataset come from three different sources – Topics, CMV, and Random

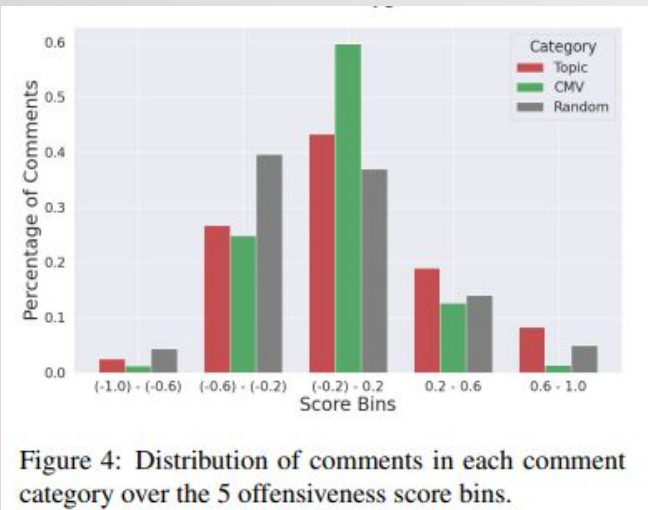


그림 14. 각 카테고리의 댓글 분포

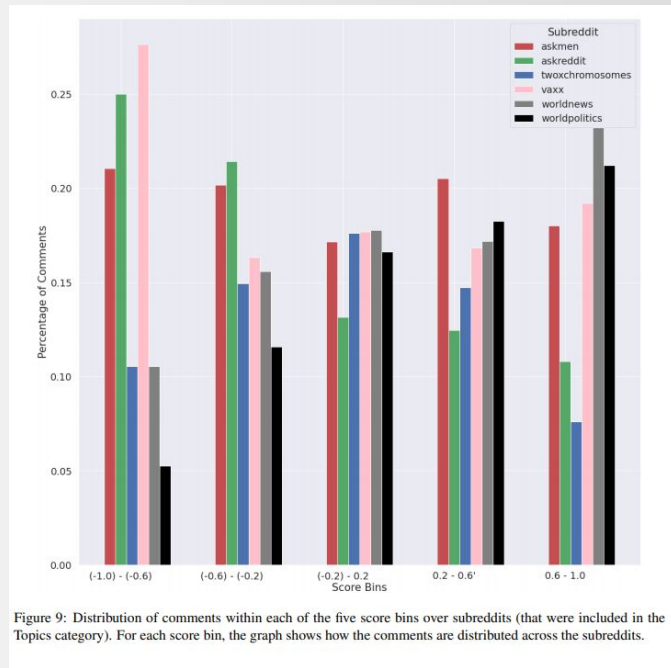


그림 15. subreddit의 댓글 분포

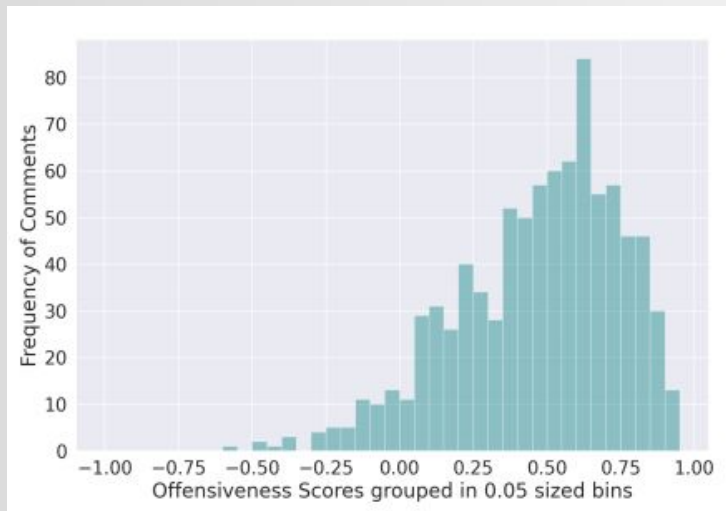


그림 16. 공격적인 단어를 포함하는 댓글의 빈도 히스토그램 (bin 0.05)

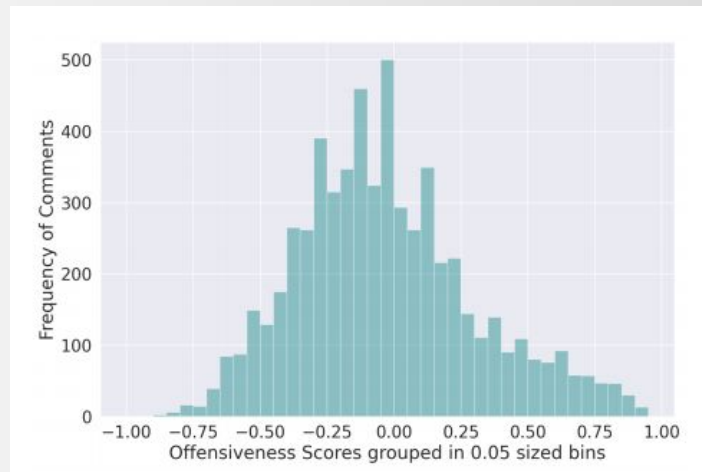


그림 10. 댓글 빈도 - 공격성 정도 (bin 0.05)

No-swearing dataset

Reduced-range

05

Computational Modeling

Bidirectional LSTM

loss function(MSE), optimizer(Adam), learning_rate(0.001),
hidden dimension(256), batch size(32), dropout(0.5), epochs(7)

BERT

loss function(MSE), optimizer(Adam), learning rate($2e - 5$),
batch size(16), epochs(3)

HateBert: BERT와 동일

Dataset	HateBERT		BERT		BiLSTM	
	r	MSE	r	MSE	r	MSE
a. Ruddit	0.886 ± 0.003	0.025 ± 0.001	0.873 ± 0.005	0.027 ± 0.001	0.831 ± 0.005	0.035 ± 0.001
b. <i>Identity-agnostic</i>	0.883 ± 0.006	0.025 ± 0.001	0.869 ± 0.007	0.027 ± 0.001	0.824 ± 0.007	0.036 ± 0.001
c. <i>No-swearing</i>	0.808 ± 0.013	0.023 ± 0.001	0.783 ± 0.012	0.027 ± 0.001	0.704 ± 0.014	0.036 ± 0.002
d. <i>Reduced-range</i>	0.781 ± 0.014	0.022 ± 0.001	0.757 ± 0.011	0.025 ± 0.001	0.659 ± 0.008	0.033 ± 0.001

Table 5: Five-fold cross-validation results of the models on Ruddit and its variants. r = Pearson's R. Note: Scores for c. and d. are not directly comparable to scores for a. and b. as they involve different score ranges.

표 5. 모델의 성능

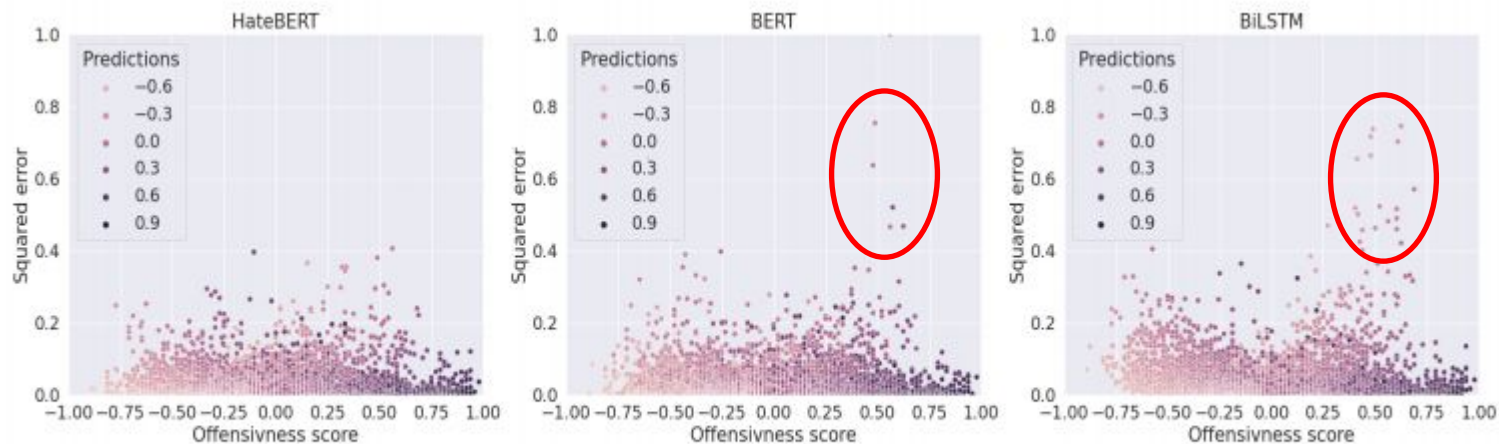


Figure 5: Squared error values for the 3 models' predictions over the offensiveness score range in Ruddit.

그림 17. 3개 모델의 오차
제공값

결론

THE

END

감사합니다
