Implicitly abusive language – What does it actually look like and why are we no getting there?

2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics

JeWoong Hwang, 2021220160

Acknowledgement & Ethical Considerations

- This paper contains real-life examples of abusive language
- The writers of the paper and I are aware that people may feel offended by these examples, and this demonstrates that implicit abuse can be extremely severe.
- These examples **in no way** reflect my opinion or the paper.

Introduction

- Abusive or offensive language
- *Hate speech* and *cyberbullying*

- The number of user-generated content is growing.
- Abusive language detection is required
  - Moderation etc.

Implicit and Explicit abusive language

- Explicit language
  - Explicit words
    - *Stop editing this, you dumbass*
    - *Go lick a pig you arab muslim piece c*

- Implicit language
  - Not conveyed by explicit words
    - *I haven't had an intelligent conversation with a woman in my whole life*
    - *Why aren't there any Mexicans on Star Trek? Because they don't work in the future either.*

Contents

- List of subtypes, and quantitative information
- Derived research tasks
- Problem of existing datasets
- Key issues of datasets

Current works and datasets

- Supervised Learning is the most prominent
  - SVM, Logistic Regression
  - They are poor at detecting implicit abusive language
- Little error analysis of system output
- Binary classification task only

dataset	publication	size	%expl.
Kumar	(Kumar et al., 2018)	15,000	32.7
SBFrames	(Sap et al., 2020)	45,318	37.6
Waseem	(Waseem and Hovy, 2016)	16,165	44.4
Warner	(Warner and Hirschberg, 2012)	3,438	51.3
OffensEval	(Zampieri et al., 2019)	13,240	54.0
Razavi	(Razavi et al., 2010)	1,525	64.7
Founta	(Founta et al., 2018)	59,357	75.9
Kaggle	(Wulczyn et al., 2017)	312,737	76.7

Table 1: Statistics of datasets (*size*: number of microposts; *expl.*: explicit abuse).

Subtypes

- Stereotypes
  - *Jews are good at making money.*
  - *Jews always support terror instability.*
  - *Jews mourn the loss of a member of their community*
  - *Jews belong in the kitchen*
  - *Women are good at making money*
- Perpetrators
  - *Our neighbor is beating his wife*

Subtypes

- Comparisons
  - *You look like an idiot*
  - *You sing like a dying bird*
  - *You look like a princess*
  - *You look like you have slept badly*
- Dehumanization
  - *A wild flock of Jews ~~~~*
  - *I own my wife and her money.*
  - *How come bunches of gay people mushroom out of the ground?*

Subtypes

- Euphemistic Constructions
  - You inspire my inner serial killer
  - I'm not excited about your existence.
- Call for Action
  - Thank you for your fortitude and perseverance. Please give McConnell a kick in the butt from some of us.
  - He should be given 5000 volts!
- Multimodal Abuse

Subtypes

- Phenomena Requiring World Knowledge and Inferences
  - Jokes, Sarcasm, Rhetorical questions, and others
    - *What's better than winning gold in the paralympics? Walking.*
    - *Welcome to the Hotel Islamfornia. You may check out any time but you can never leave.*

Distribution

	datasets					
subtype	**Kumar**	**SBFrames**	**Waseem**	**Warner**	**OffensEval**	***average***
other implicit abuse	9.8	28.4	12.8	30.4	2.4	16.8
perpetrator	18.2	2.4	22.0	17.1	15.2	15.0
stereotype	13.4	2.0	12.2	20.0	14.2	12.4
joke	0.0	40.8	0.2	2.5	0.0	8.7
call for action	3.8	1.6	1.0	4.6	2.8	2.8
dehumanization	2.2	0.6	1.0	2.5	3.0	1.9
euphemistic construction	1.4	0.6	2.0	1.3	3.8	1.8
rhetorical question	1.2	1.6	1.6	2.1	0.6	1.4
comparison	0.6	0.0	1.4	0.0	0.0	0.4
sarcasm	1.0	0.0	0.2	0.0	0.6	0.4
unknown	37.0	11.0	37.8	10.8	23.0	23.9
explicit abuse (abus. word missing in Wiegand et al. (2018))	11.4	10.0	7.8	8.8	34.4	14.5

Table 2: Percentage of different subtypes of implicit abuse (including overlooked explicit abuse) within a dataset. The numbers are obtained by manually inspecting 500 implicit texts from each of the datasets.

What should datasets look like? : Bias

- Model might suffer from bias
  - Identity group (e.g. Jews, Muslims)
  - Jokes from *reddit.com*
  - *I'm pretty sure Hitler just said "I wanna glass of juice" not I wanna gas the <IDENTITY_GROUP>.*
  - *Being a <IDENTITY_GROUP> I have a confusion choosing my career. Either to go with ISIS or Al-Qaeda?*

What should datasets look like?: Negative

Lack of negative data.

Divide & conquer

dataset	publication	size	%expl.
Kumar	(Kumar et al., 2018)	15,000	32.7
SBFrames	(Sap et al., 2020)	45,318	37.6
Waseem	(Waseem and Hovy, 2016)	16,165	44.4
Warner	(Warner and Hirschberg, 2012)	3,438	51.3
OffensEval	(Zampieri et al., 2019)	13,240	54.0
Razavi	(Razavi et al., 2010)	1,525	64.7
Founta	(Founta et al., 2018)	59,357	75.9
Kaggle	(Wulczyn et al., 2017)	312,737	76.7

Table 1: Statistics of datasets (*size*: number of microposts; *expl.*: explicit abuse).

identity group	woman	lesbian	gay	black	muslim	jew
% abusive	67.3	71.7	75.2	87.2	87.8	93.8

Table 3: Abusive posts with identity group.

	datasets					
subtype	Kumar	SBFrames	Waseem	Warner	OffensEval	*average*
other implicit abuse	9.8	28.4	12.8	30.4	2.4	16.8
perpetrator	18.2	2.4	22.0	17.1	15.2	15.0
stereotype	13.4	2.0	12.2	20.0	14.2	12.4
joke	0.0	40.8	0.2	2.5	0.0	8.7
call for action	3.8	1.6	1.0	4.6	2.8	2.8
dehumanization	2.2	0.6	1.0	2.5	3.0	1.9
euphemistic construction	1.4	0.6	2.0	1.3	3.8	1.8
rhetorical question	1.2	1.6	1.6	2.1	0.6	1.4
comparison	0.6	0.0	1.4	0.0	0.0	0.4
sarcasm	1.0	0.0	0.2	0.0	0.6	0.4
unknown	37.0	11.0	37.8	10.8	23.0	23.9
explicit abuse (abus. word missing in Wiegand et al. (2018))	11.4	10.0	7.8	8.8	34.4	14.5

Table 2: Percentage of different subtypes of implicit abuse (including overlooked explicit abuse) within a dataset. The numbers are obtained by manually inspecting 500 implicit texts from each of the datasets.

Classification below the Micropost-level

- Mentions of abuse
  - *@USER exposes the hypocrisy of claims that [Muslims want to suppress free speech].*
  - *The Texas GOP thinks that [gay people need a cure]*
- Task to isolate abusive clauses

Role of machine learning

- Sophisticated model < quality dataset
  - *Asian children are intelligent.*
  - *All Asian people lie.*
- Supervised network?

Conclusion

- Subtypes
- Dataset and bias
  - Negative datas

Thank you