university of groningen

NLP Against Toxicity

Detecting Toxic Interactions in Social Media

Tommaso Caselli
GroNLP, University of Groningen - Jantina Tammes School
t.caselli@rug.nl

Overview

- Toxicity: How many language phenomena?
- Abusive Language: Data and Dataset creation
- Abusive Language Detection
  - Model specific solutions: HateBERT
- Take-home message

Social Media

The Big Promise

- increase connectivity
- a more open world
- productive engagements
- diversity of opinions/perspectives
- “collective intelligence”

https://lifedisrobed.files.wordpress.com/2015/09/social-media-banner1.png

Social Media

Something something something something … Dark Side

- abusive language
- cyberbullying
- trolling
- hate speech
- personal attacks
- toxic interactions

What is Toxicity?

Unfortunately not this

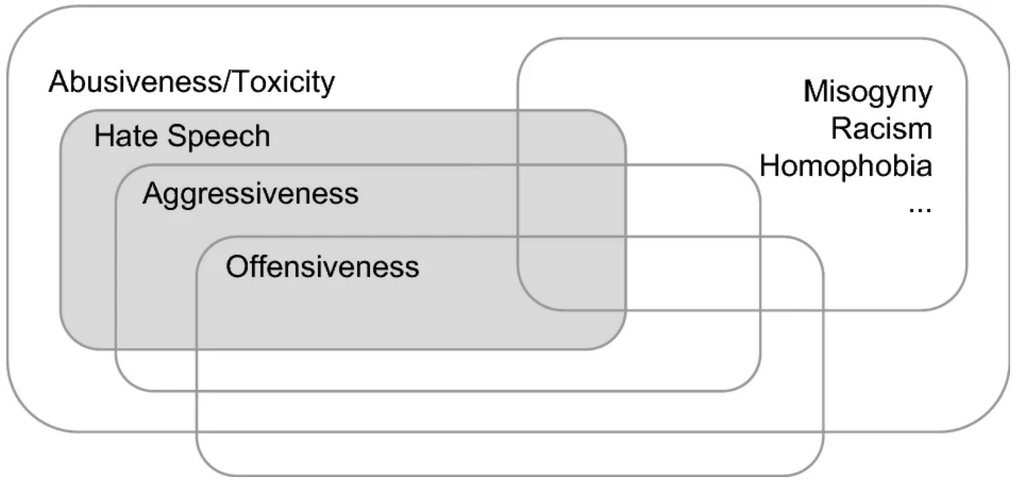
Conversion, software version 7.0
Looking at life through the eyes of a tire hub
Eating seeds as a pastime activity
The toxicity of our city, our city

You, what do you own the world?
How do you own disorder? Disorder
Now somewhere between the sacred silence
Sacred silence and sleep
Somewhere, between the sacred silence and sleep
Disorder, disorder, disorder

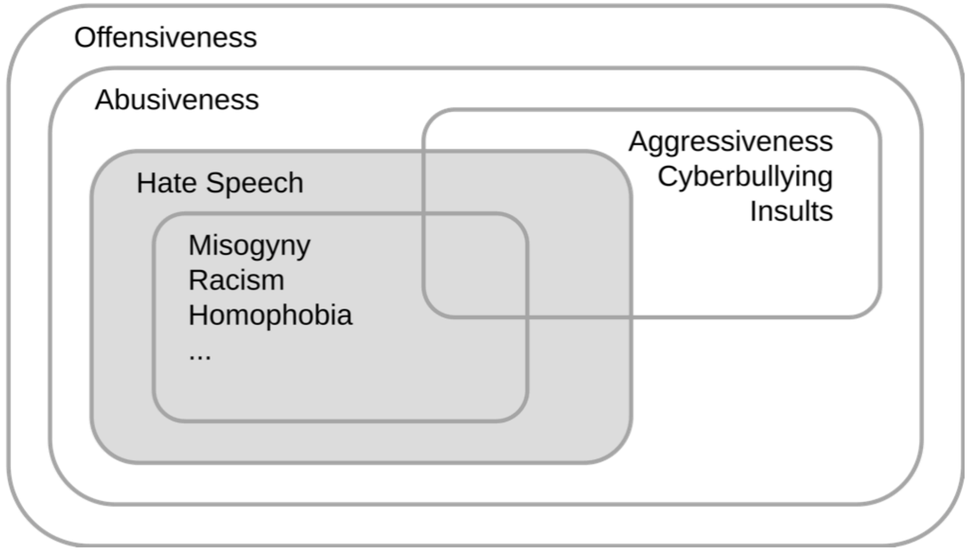
What is Toxicity?

SOCIAL UNACCEPTABLE LANGUAGE

- abusive language
- cyberbullying
- trolling
- hate speech
- insults
- personal attacks
- fear speech



https://link.springer.com/article/10.1007/s10579-020-09502-8



https://aclanthology.org/2020.peoples-1.15/

H. Van der Veen - IK Master Thesis 2022

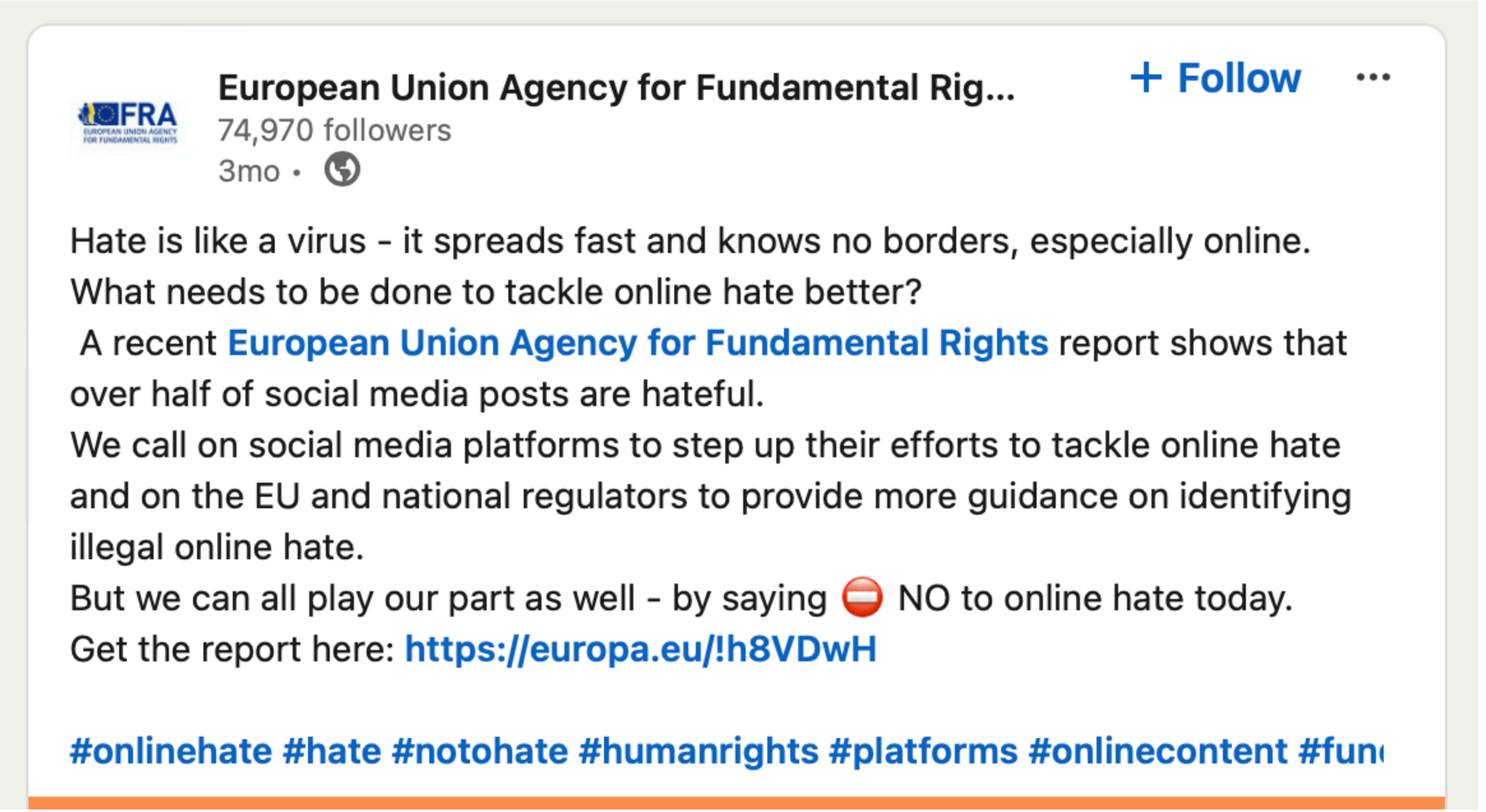
“Bad Language” and the Myth of Freedom of Speech

- Use of “bad language” is intimately connected with freedom of speech
- Free speech must have a rational end: facilitate communication between citizens
- Free speech ≠ Absolute speech
- U.S. Supreme Court “has not protected […] libel, slander, perjury, false advertising, obscenity and profanity, solicitation of a crime, or ‘fighting' words” [Francis Caravan - *Freedom of Expression: Purpose As Limit*]

Things are serious …

- Pew Research Center (2017) - U.S.A.:
  - 41% personally experienced harassment online
  - 66% witnessed harassment directly towards others
- 22% experienced offensive name-calling
- 10% threats

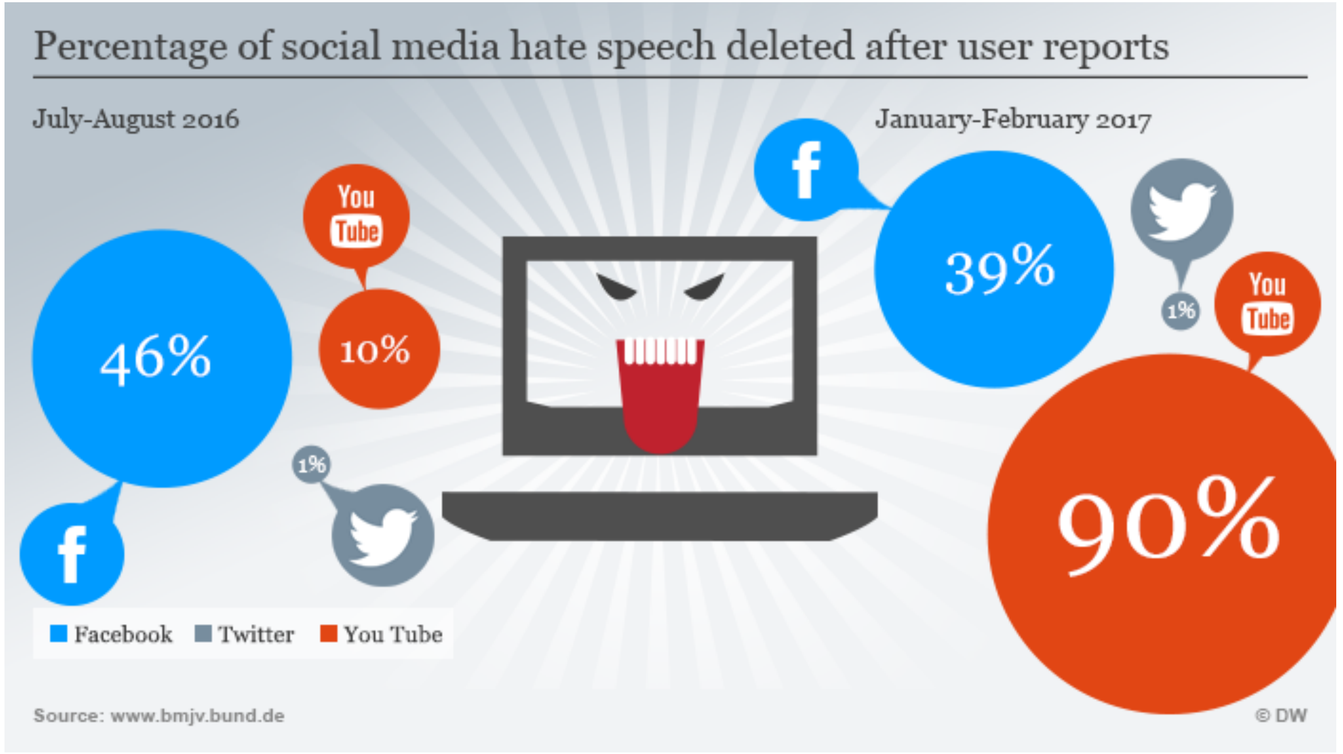
Things are serious …



https://www.linkedin.com/posts/eu-fundamental-rights-agency_onlinehate-hate-notohate-activity-7158730719148535810-gKUd/

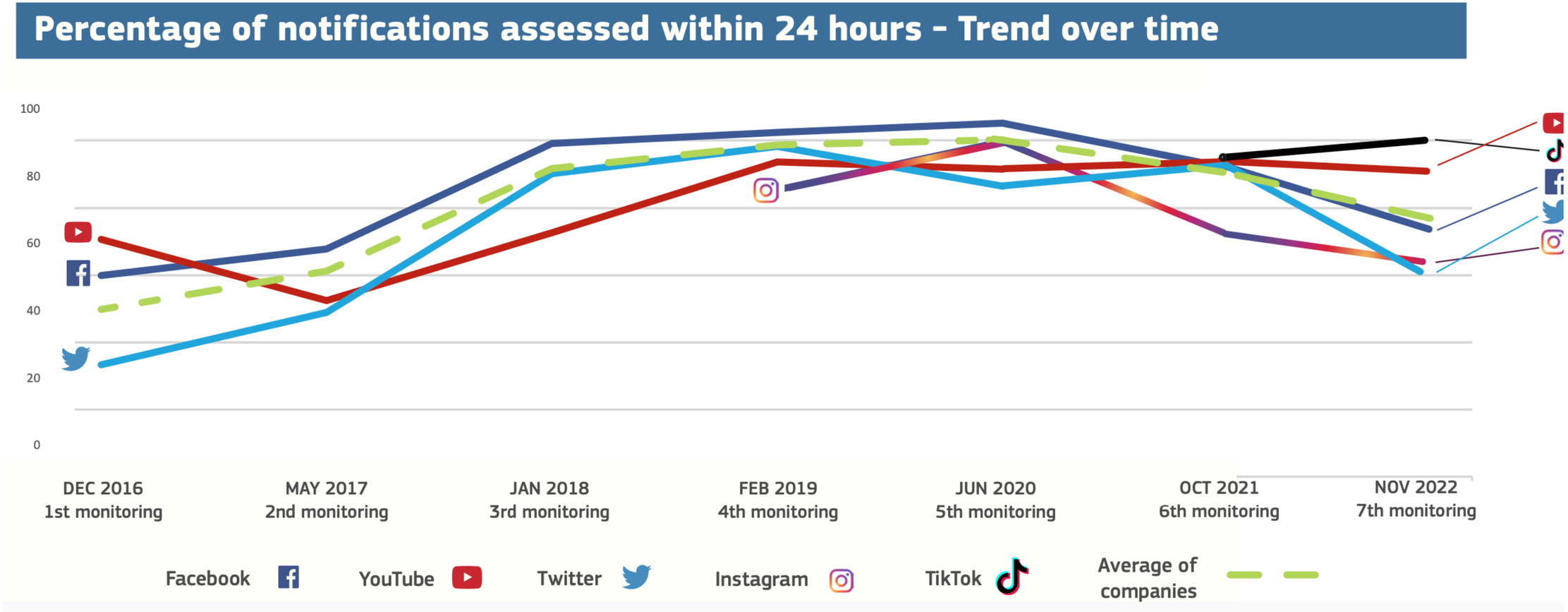
… some measures have been taken

- Social Media companies have agreed with EU Commission in 2016 to a code of Conduct on illegal online hate speech.
- Germany passed a law in 2017 that imposes Social Media companies to swiftly remove online hateful content



https://m.dw.com/image/37928737_105.png

… but we are slowing down



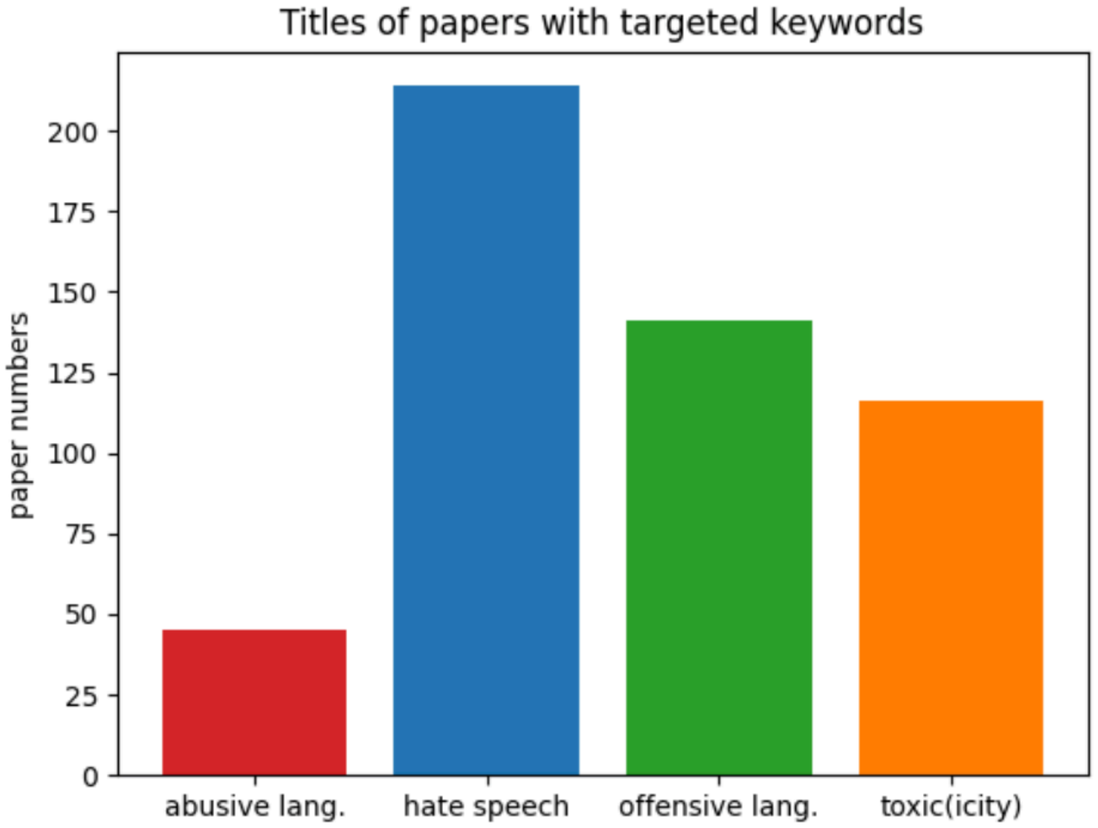
Overall, IT companies removed **63.6%** of the content notified to them, while **36.4%** remained online. This result is slightly higher than the average of **62.5%** recorded in 2021, but lower than the peak score of **71%** in 2020.

https://commission.europa.eu/system/files/2022-12/Factsheet - 7th monitoring round of the Code of Conduct.pdf

Abusive Language and NLP

A trendy topic

- ACL Anthology (2020 / 2024) : 32.475 papers
- Papers dedicated to Socially Unacceptable Language: 1.6%



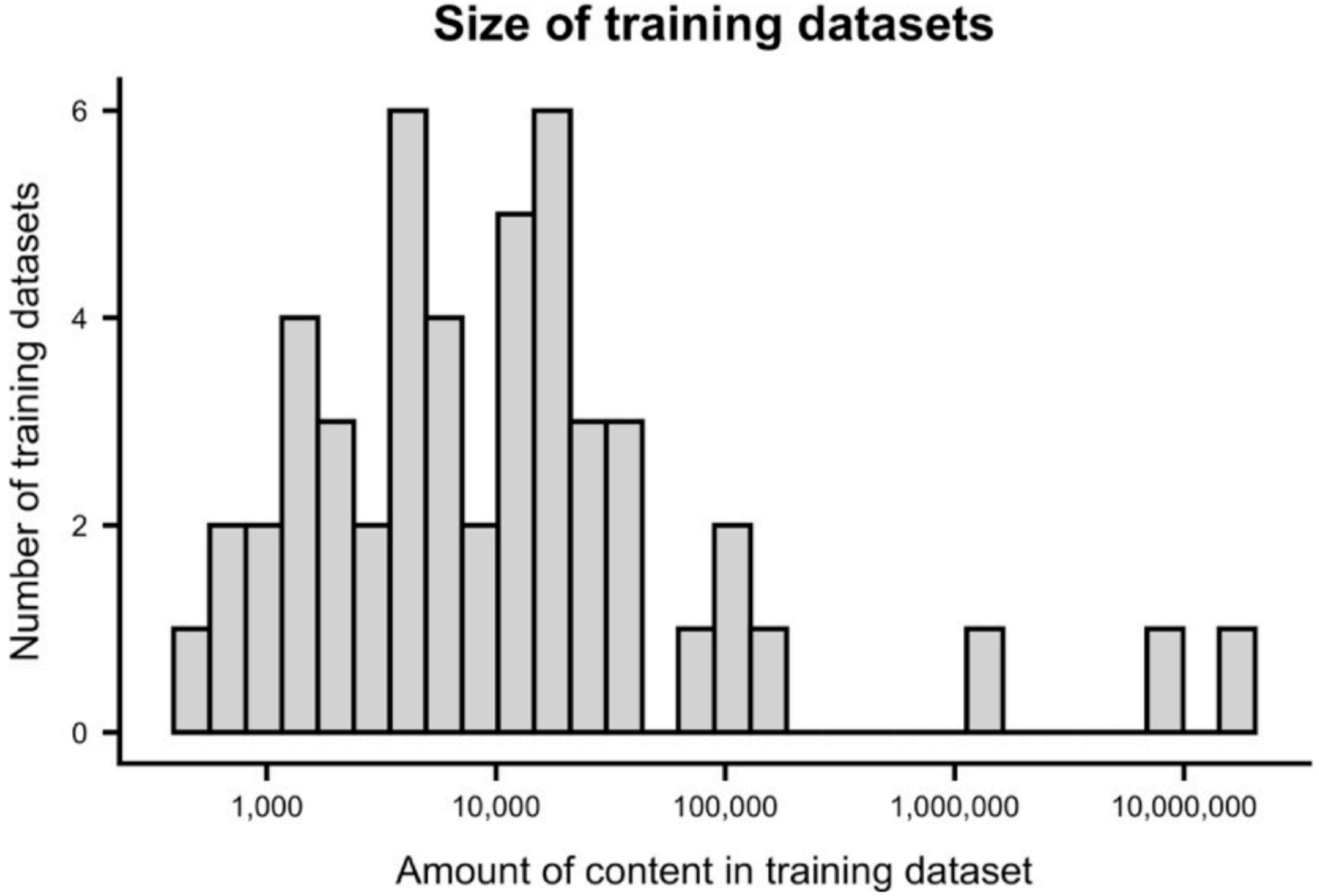
Abusive Language and NLP

Some updated figures

- 25 documented languages (2024)
  - +13 from 2020 (108% increase!)
  - Twitter/X is still the most common platform (53 datasets)
  - Other 13 datasets mixed (Twitter/X + Gab)

How many data?

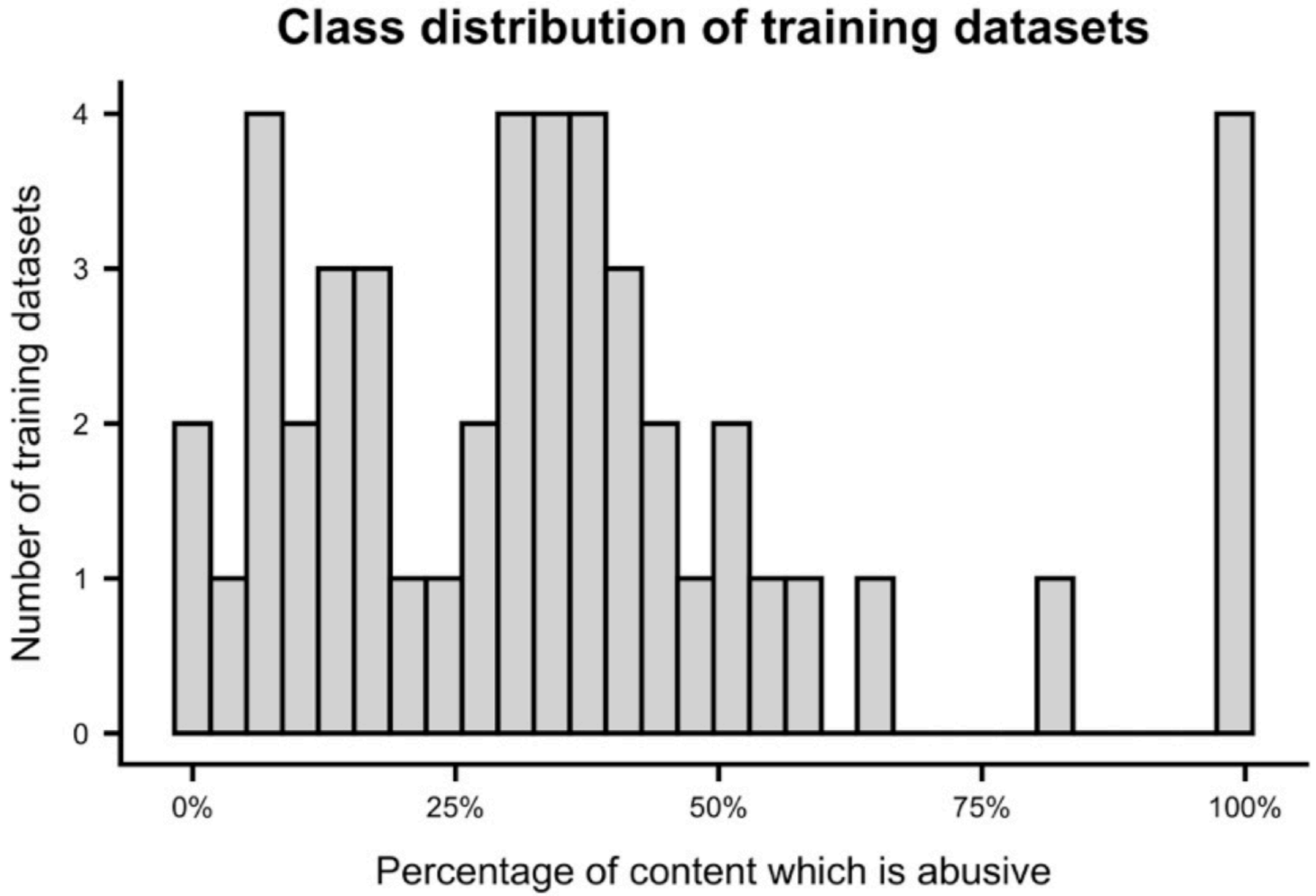
Sizes of current datasets - training only



Vidgen & Derczynski 2021

How much “abuse”

Distribution of the positive class in existing training data



Vidgen & Derczynski 2021

What are we talking about ?

- What is Abusive Language?
- What is Offensive Language?
- What is Hate Speech?
- What is Toxic Language?
- What is Cyberbullying?

Guess the label

- *My boyfriend sends me the sweetest messages all the time and my first reaction is "What the fuck???”* [Founta et al., 2018]
  - A: Abusive
  - B: Offensive
  - C: Hateful
  - D: None

Guess the label

- *@FoxNews @BarackObama Like he ever kept out any threats. He's lying as usual. #BuildThatWall* [SemEval 2019: Task 5 HatEval]
  - A: Abusive
  - B: Offensive
  - C: Hateful
  - D: None

Guess the label

- *My boyfriend sends me the sweetest messages all the time and my first reaction is "What the fuck???”* [Founta et al., 2018] - **ABUSIVE**
- *@FoxNews @BarackObama Like he ever kept out any threats. He's lying as usual. #BuildThatWall* [SemEval 2019: Task 5 HatEval] - **NONE**

Trendy but with problems

- Datasets tend to be made available each with its own definition of the phenomenon
- Negative aspects:
  - confusion
  - comparison (lack of)
    - data & systems

What is offensive?

	OFFENSIVE LANGUAGE
OXFORD DICTIONARY	1. Causing someone to feel resentful, upset, or annoyed. 2. [attributive] Actively aggressive; attacking.
Founta et al., 2018	Profanity, strongly impolite, rude or vulgar language expressed with fighting or hurtful words in order to insult a targeted individual or group. (Chen et al. 2012), (Razavi et al. 2010).
Zampieri et al., 2019	[W]e label a post as offensive (OFF) if it contains any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.

What is abusive?

	ABUSIVE LANGUAGE
OXFORD DICTIONARY	1. Extremely offensive and insulting. 2. Engaging in or characterized by habitual violence and cruelty.
Founta et al., 2018	Any strongly impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion. (Papegnies et al. 2017), (Park and Fung 2017), (Nobata et al. 2016).
Fortuna and Nunes, 2018	The term abusive language was used to refer to hurtful language and includes hate speech, derogatory language and also profanity.

What is hateful?

	HATEFUL LANGUAGE / HATE SPEECH
OXFORD DICTIONARY	Abusive or threatening speech or writing that expresses prejudice against a particular group, especially on the basis of race, religion, or sexual orientation.
Founta et al., 2018	Language used to express hatred towards a targeted individual or group, or is intended to be derogatory, to humiliate, or to insult the members of the group, on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. (Davidson et al. 2017), (Badjatiya et al. 2017), (Warner and Hirschberg 2012), (Schmidt and Wiegand 2017), (Djuric et al. 2015).
Sanguinetti et al., 2017	that is abusive, insulting, intimidating, harassing, and/or incites to violence, hatred, or discrimination. It is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth (Erjavec and Kovacič, 2012).
Mondal et al., 2017	An offensive post, motivated, in whole or in a part, by the writer's bias against an aspect of a group of people

Different definitions, different dimensions

- Language:
  - “impolite, rude or hurtful language”
  - “offensive (OFF) if it contains”
- Intentions:
  - “hatred towards a targeted individual or group”
  - “used to express hatred”
  - “is intended to be derogatory, to humiliate, or to insult”

Different definitions, different dimensions

- Effects:
  - “show a debasement of someone or something”
  - “[c]ausing someone to feel resentful, upset, or annoyed”
- Targets:
  - “a targeted individual or group”
  - “[i]t is directed against people on the basis of their race, ethnic origin, religion, gender, age, physical condition, disability, sexual orientation, political conviction, and so forth”

How to recognise abusive language?

Annotating data

- Waseem et al;., 2017: proposal of a **hierarchical annotation**
- better distinction of different phenomena and subtasks
- **Hierarchical:**
  - distinction of dimensions
  - separate annotations of each dimension

What is needed to annotate abusive language?

- **Definition**
  - must address an appropriate level of “abstraction”
  - not too generic (“annotate abuse”), nor too narrow
    - “hate denotes a specific aggressive and emotional behavior, excluding other varieties of abuse, such as dismissal, insult, mistrust and belittling.” Vidgen et al., 2019
  - it should assume little about speakers’ intentions; an unsuitable basis for definitions of abuse

What is needed to annotate abusive language?

- **Target**
  - towards who is directed the abuse?
    - individual, group of individuals (e.g. members)
    - concept:
      - *Capitalism sucks* vs. *Islam sucks*
    - a generic target

What is needed to annotate abusive language?

- **Explicitness**
  - How explicit/overt is the abuse?
  - Distinction between ***denotation*** *(*the literal meaning of a term or symbol), and ***connotation*** (cultural or emotional associated meaning of a term or symbol)
    - *inexpensive vs cheap ; slim vs skinny; rose*

What is needed to annotate abusive language?

- Explicit: “language which is unambiguous in its potential to be abusive, for example language that contains racial or homophobic slurs.” Waseem et al. 2017

- Implicit: “Implicit abusive language is that which does not immediately imply or denote abuse […] use of ambiguous terms, sarcasm, lack of profanity or hateful terms, and other means, generally making it more difficult to detect” Waseem et al. 2017

.

Typology of Abuse

Waseem et al., 2017

	Explicit	*Implicit*
Directed	"Go kill yourself", "You're a sad little f*ck" (Van Hee et al., 2015a), "@User shut yo beaner ass up sp*c and hop your f*ggot ass back across the border little n*gga" (Davidson et al., 2017), "Youre one of the ugliest b*tches Ive ever fucking seen" (Kontostathis et al., 2013).	"Hey Brendan, you look gorgeous today. What beauty salon did you visit?" (Dinakar et al., 2012), "((((@User)))) and what is your job? Writing cuck articles and slurping Google balls? #Dumbgoogles" (Hine et al., 2017), "you're intelligence is so breathtaking!!!!!!" (Dinakar et al., 2011)
Generalized	"I am surprised they reported on this crap who cares about another dead n*gger?", "300 missiles are cool! Love to see um launched into Tel Aviv! Kill all the g*ys there!" (Nobata et al., 2016), "So an 11 year old n*gger girl killed herself over my tweets? ^_^ thats another n*gger off the streets!!" (Kwok and Wang, 2013).	"Totally fed up with the way this country has turned into a haven for terrorists. Send them all back home." (Burnap and Williams, 2015), "most of them come north and are good at just mowing lawns" (Dinakar et al., 2011), "Gas the skypes" (Magu et al., 2017)

How to create a reliable datasets for abusive language?

An not-so-easy task

- Abusive language is actually sparse
  - Founta et al., 2018’s estimate: between 0.1% and 3% of the messages in Twitter
  - Saha et al., 2023 identified 700k hate speech post in a collection of 21M posts from Gab —> 3.3%
- Random sampling not a feasible solution

How to create a reliable datasets for abusive language?

Available strategies

- Three approaches:
  - use of communities;
  - use of keywords representing targets of abuse (e.g. women)
  - use of potentially hateful/abusive users

How to create a reliable datasets for abusive language?

Bias in data

- Keywords (or users) can add bias in data
  - Waseem and Hovy (2016) extract tweets matching query words likely to co-occur with abusive content
- Bias can give rise to “exclusion” (Hovy & Spruit, 2016)
  - majority of cases Afro-american dialect messages are marked as abusive

rank	Founta	Waseem
1	bitch	**commentator**
2	niggas	comedian
3	motherfucker	**football**
4	fucking	**announcer**
5	nigga	pedophile
6	idiot	mankind
7	asshole	sexist
8	fuck	**sport**
9	fuckin	outlaw
10	pussy	driver

Wiegand et al., 2019 - https://www.aclweb.org/anthology/N19-1060.pdf

DALC: Dutch Abusive Language Corpus

Language phenomena & definitions

- Comprehensive and unique resource for (generic) **abusive** and **offensive** language in Dutch in Twitter

Offensive Language (Zampieri et al., 2019a)	**Abusive Language** (Caselli et al., 2021)
Posts containing any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct. This includes insults, threats, and posts containing profane language or swear words.	Impolite, harsh, or hurtful language (that may contain profanities or vulgar language) that result in a debasement, harassment, threat, or aggression of an individual or a (social) group, but not necessarily of an entity, an institution, an organisations, or a concept.

DALC: Dutch Abusive Language Corpus

Language phenomena: differences

- Strictly connected but inherently different phenomena

Offensive Language	Abusive Language
Highly subjective	It can be detailed (i.e., parameters to identify abusive expressions)
Depend on the context of occurrence	Depend (mainly) on the content
Focused on the effects	It takes into account intentions and effects
Presence of a target is optional	A target must always be present

DALC: Dutch Abusive Language Corpus

Putting different data collection methods together

- RUG Twitter Corpus (Tjong Kim Sang, 2011)
  - Keywords: TF-IDF from controversial posts  reddit `r/thenetherlands` (van Rosendaal et al., 2020)
  - Geolocation: economically depressed areas in the Netherlands
  - Seed users (Wiegand et al., 2018; Ribeiro et al., 2018)

DALC: Dutch Abusive Language Corpus

Hierarchical annotation: explicitness

- Explicitness layer (Waseem et al., 2017; Zampieri et al., 2019)
  - *Surface evidence*
  - *Intentions of the producers*
  - *Effect on the receivers*
- Core values:
  - **Explicit**: ABU/OFF + *contains a profanity or a slur*
  - **Implicit**: ABU/OFF + *no profanity or a slur*

DALC: Dutch Abusive Language Corpus

Hierarchical annotation: target

- Target layer (Waseem et al., 2017; Zampieri et al., 2019)
  - *to whom the message is directed (to)*
- 4 values - same as in Zampieri et al., 2019:
  - **Individual**: a person, named or unnamed
  - **Group**: a group of individuals considered as a unity
  - **Other**: organizations; institutions; concepts
  - **Not**: offensive but no target

DALC: Dutch Abusive Language Corpus

Target annotation

TEXT	ABU	OFF	TGT
@USER OMDAT VROUWEN MOEILIJKE WEZENS ZIJN (buik van vol) (*@USER because women are difficult creatures (belly full of them*)	NOT	IMPLICIT	GROUP
@USER @USER @USER @USER Fucking clown, zwarte piet word racistisch ERVAREN ! en dat komt slechts door onwetenheid (*@USER @USER @USER @USER Fucking clown Black Piet is being racist EXPERIENCE ! and that is only because of ignorance*)	EXPLICIT	EXPLICIT	IND.
Er is een fucking mug groot als mijn duim (*There is a fucking mosquito as big as my thumb)*	NOT	EXPLICIT	NOT
@USER @USER Goed zo @USER dat doe je wijs,ze komt jou even vertellen wat je moet doen nadat ze je vals beschuldigd,ik zou ze ook niet moeten. (*@USER @USER Good @USER you are doing that wisely,she comes to tell you what to do after she accuses you falsely,I wouldn't like them either*.)	NOT	NOT	—

DALC: Dutch Abusive Language Corpus

Annotation process

- 4 annotation rounds
  - parallel annotations for OFF
  - 11,292 tweets
  - No overlap TRAIN / TEST
  - EXPLICIT instances are majority
  - 6.95% OFF has no target

Annotated Dimension	Subclass	Train	Dev	Test	Total
Abusive	EXP	855	127	328	1,310
	IMP	536	116	135	787
	NOT	5,426	962	2,807	9,195
Offensive	EXP	1,407	230	584	2,221
	IMP	1,070	209	283	1,562
	NOT	4,340	766	2,403	7,509
Target - Abusive	IND	777	127	254	1,158
	GRP	470	87	158	715
	OTH	144	29	51	224
Target - Offensive	IND	1,147	191	361	1,699
	GRP	705	133	244	1,082
	OTH	489	93	157	739
	NOT	136	22	105	263

DALC: Dutch Abusive Language Corpus

Annotated data & possible experiment settings

- The multi-layered annotation format allows for multiple classification experiments

PHENOMENON	COARSE-GRAINED	FINE-GRAINED
OFFENSIVE LANGUAGE	2 classes	3 classes
ABUSIVE LANGUAGE	2 classes	3 classes
OFFENSIVE vs ABUSIVE	3 classes	X
TARGET	X	3/4 classes

We happy?

We happy?

Not really

- Language is dynamic - it continually evolves
- Trained models are subject to potential bias in their training data
- We want models to *generalize well*
  - across data —> model the phenomenon not the dataset
  - over time —> in 2 year time, can we detect ABU messages?

Dynamic & Functional Benchmarks

OP-NL & HateCheck-NL

OP-NL & HateCheck-NL

Dynamic vs. Functional

- Dynamic benchmarks: data annotated with same annotation guidelines, **but** from a different data distribution
  - Different time period / same medium
  - Different medium
- Functional benchmarks: systematically generated test cases aiming at evaluating in a task-agnostic methodology trained models
  - One makes a list of test cases that targets generalisation functionalities of models
  - Examples are somewhat independent from a specific task

Dynamic Benchmark: OP-NL

- OP- NL contains 1,500 tweets from March 2021 with at least one mention of a Dutch politician from *Tweede Kamer*
- Same definition of offensive language as DALC:
  - "*messages containing any form of non-acceptable language (profanity) or a targeted offence, which can be veiled or direct" (Zampieri et al., 2019)*
- 961 messages (64%) NOT vs. 539 messages (36%) OFF

Functional Benchmark: HateCheck-NL

- Adaptation to Dutch of HateCheck (Röttger et al., 2021)
  - **hate**: "*abuse that is targeted at a protected group or at its members for being a part of that group*."
  - **compatible** with DALC Abusive language
- 3,835 functional tests instances across the 29 functionalities:
  - 18 functionalities for expressions of hate
  - 11 functionalities for non-hateful expressions

Functional Benchmark: HateCheck-NL

Functionalities

	Functionality	Description from Röttger et al. (2021)	Label	Count	
				templ	cases
F1	derog_neg_emote_h	Strong negative emotions explicitly expressed about a protected group or its members	hateful	20	140
F2	derog_neg_attrib_h	Explicit descriptions of a protected group or its members using very negative attributes	hateful	20	140
F3	derog_dehum_h	Explicit dehumanisation of a protected group or its members	hateful	20	140
F4	derog_impl_h	Implicit derogation of a protected group or its members	hateful	20	140
F5	threat_dir_h	Direct threats against a protected group or its members	hateful	20	140
F6	threat_norm_h	Threats expressed as normative statements	hateful	20	140
F7	slur_h	Hate expressed using slurs	hateful	10	170
F8	slur_homonym_nh	Non-hateful homonyms of slurs	non-hate	25	25
F9	slur_reclaimed_nh	Use of reclaimed slurs	non-hate	45	45
F10	profanity_h	Hate expressed using profanity	hateful	20	140
F11	profanity_nh	Non-hateful uses of profanity	non-hate	100	100
F12	ref_subs_clause_h	Hate expressed through pronoun reference in subsequent clauses	hateful	20	140
F13	ref_subs_sent_h	Hate expressed through pronoun reference in subsequent sentences	hateful	20	140
F14	negate_pos_h	Hate expressed using negated positive statements	hateful	20	140
F15	negate_neg_nh	Non-hate expressed using negated hateful statements	non-hate	20	140

Functional Benchmark: HateCheck-NL

How to?

- Examples are generated using templates:
  - Fixed form
  - Many possible fillers
  - Easy to manipulate
- `I hate [IDENTITY_TERM]`
- `[IDENTITY_TERM]`: gay, trans people, women, Muslims, immigrants, …

Functional Benchmark: HateCheck-NL

Examples

```
derog_neg_emote_h,4,Ik haat zwarte personen.,hateful,black people,,,1,Ik haat [IDENT_P].
derog_neg_emote_h,5,Ik haat gehandicapten.,hateful,disabled people,,,1,Ik haat [IDENT_P].
derog_neg_emote_h,6,Ik haat immigranten.,hateful,immigrants,,,1,Ik haat [IDENT_P].
derog_neg_emote_h,7,Ik haat Moslims.,hateful,Muslims,,,1,Ik haat [IDENT_P].
```

```
slur_h,892,Nog zo'n teef.,hateful,women,,,124,Nog zo'n [SLR_S].
slur_h,893,Nog zo'n slet.,hateful,women,,,124,Nog zo'n [SLR_S].
slur_h,894,Nog zo'n hoer.,hateful,women,,,124,Nog zo'n [SLR_S].
```

```
slur_homonym_nh,1011,Ik krijg deze pot niet open.,non-hateful,,,,131,Ik krijg deze pot niet open.
slur_homonym_nh,1012,Kun je mij die pot pindakaas geven?,non-hateful,,,,132,Kun je mij die pot pindakaas geven?
```

Pending Issues

The missing piece: CONTEXT

- All messages are annotated in isolation
- Communication happens in context
  - *@USER Bro go watch how he hold his backwood in the interview like that shit was irritating lmao nigga is mad suspicious* [OLID 60130]
- Abuse is sensitive to the **context** and **community** of occurrence

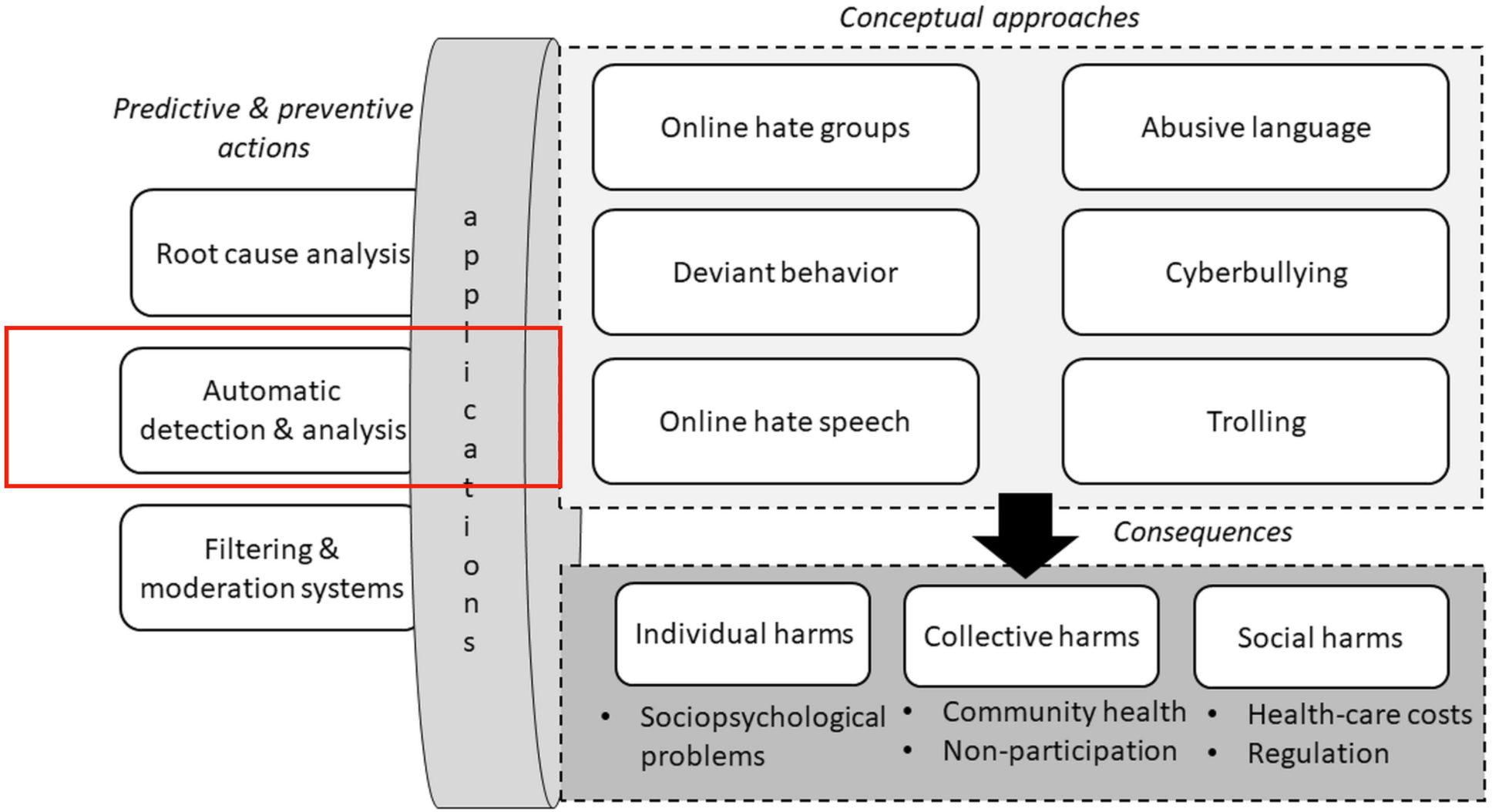
Pending Issues

Agreement

- Subjective task
- Agreement among annotators vary
- Good definitions and training of annotators are essential
- Knowing who is annotating the data is very important:
  - crowd?
  - experts?

— BREAK —

A framework for Online Abusive Language Research



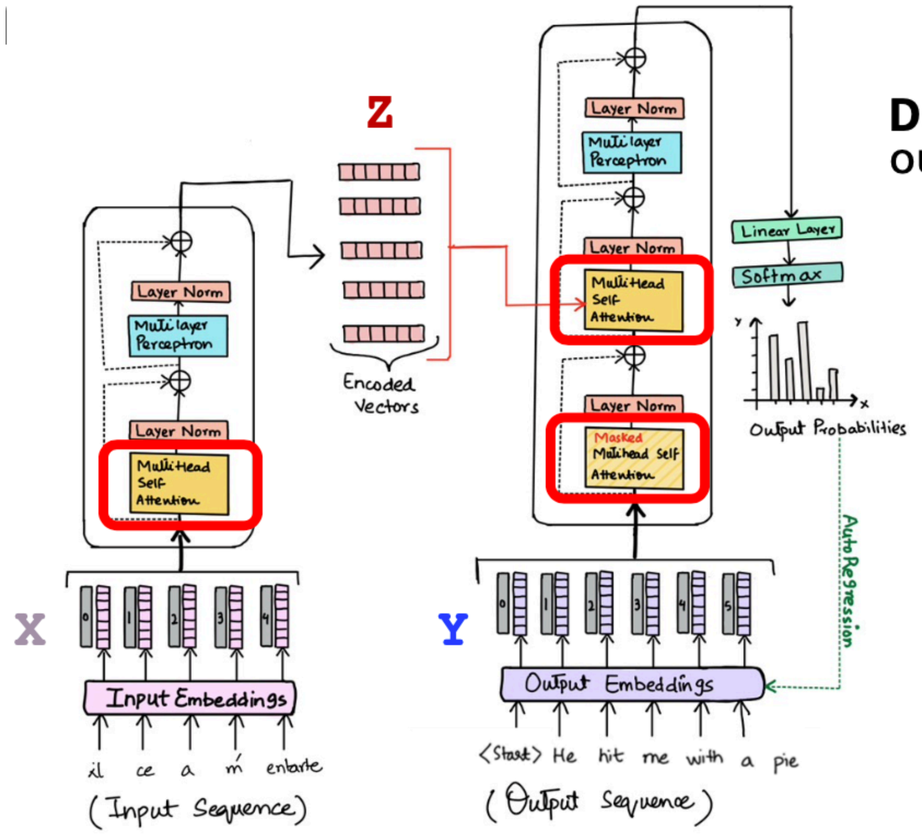
Salminen et al., 2020 - https://hcis-journal.springeropen.com/articles/10.1186/s13673-019-0205-6/figures/1

https://www.deviantart.com/iky92791/art/Welcome-to-the-Team-340596699

NLP with Transformers

Basic architecture

Encoder: Maps input sequence $x = (x_1, \ldots, x_n)$ to continuous representations $z = (z_1, \ldots, z_n)$.



Decoder: Decoder uses z to generate output sequence $y = (y_1, \ldots, y_m)$

Source Croce and Hromei 2023: https://github.com/crux82/CLiC-it_2023_tutorial

Transformers in Action

Experiments with DALC

- Encoder-based, monolingual model: BERTje  (de Vries et al., 2019)
  - `case-based` 12 layers
- Minimal pre-processing:
  - convert emoji to text
  - replace numbers [NUMBER]
  - remove # from hash-tags
- Fine-tuning + linear classifier on top

Tranformers in Action

DALC: Offensive Langauge

- Binary / coarse-grained

System	Class	Precision	Recall	Macro-F1
Dummy	OFF	0.0	0.0	0.423
	NOT	0.734	1.0	
SVM	OFF	0.644	0.513	0.718
	NOT	0.836	0.898	
BERTje	OFF	0.721	0.693	**0.802**
	NOT	0.891	0.903	

- Ternary / fine-grained

System	Class	Precision	Recall	Macro-F1
SVM	EXP	0.710	0.395	0.543
	IMP	0.297	0.212	
	NOT	0.820	0.936	
BERTje	EXP	0.762	0.639	**0.663**
	IMP	0.374	0.434	
	NOT	0.887	0.904	

- Binary classification much easier
- Difficulties between IMPLICIT and NOT

Tranformers in Action

DALC: Abusive Langauge

- Binary / coarse-grained

System	Class	Precision	Recall	Macro-F1
Dummy	ABU	0	0	0.399
	NOT	0.664	1.0	
SVM	ABU	0.858	0.323	0.655
	NOT	0.740	0.973	
BERTje	ABU	0.850	0.500	**0.748**
	NOT	0.791	0.955	

- Ternary / fine-grained

System	Class	Precision	Recall	Macro-F1
SVM	EXP	0.805	0.270	0.433
	IMP	0.461	0.033	
	NOT	0.719	0.986	
BERTje	EXP	0.759	0.447	**0.561**
	IMP	0.373	0.189	
	NOT	0.790	0.962	

- Offensive easier than Abusive
- IMPLICIT Abuse is very difficult (issues in training instances)

Tranformers in Action

DALC: Offensive vs Abusive

- Simple SVM highly competitive
- ABU is confused with OFF
- OFF is confused with NOT
- Tendency to assign ABU class to EXPLICIT messages (OFF or ABU)

Model	Class	P	R	Macro-F1
Dummy	OFF	0.0	0.0	0.2824
	ABU	0.0	0.0	
	NOT	0.7348	1.0	
SVM	OFF	0.3588	0.1510	0.5206
	ABU	0.5387	0.4212	
	NOT	0.8229	0.9376	
BERTje	OFF	0.3301	0.3391	**0.5890**
	ABU	0.5696	0.5011	
	NOT	0.8971	0.9800	

Tranformers in Action

DALC: Portability of trained models

System	Class	Precision	Recall	Macro-F1
Dummy	OFF	0.0	0.0	0.423
	NOT	0.734	1.0	
SVM	OFF	0.644	0.513	0.718
	NOT	0.836	0.898	
BERTje	OFF	0.721	0.693	**0.802**
	NOT	0.891	0.903	

	OP-NL
Model	**Macro-F1**
Dummy	39.04
BERTje OFF	**73.56**
BERTje ABU	57.57
BERTje OFF-ABU	71.85

- BERTje OFF model best performing
- Lower scores (not directly comparable) but ok
- DALC is a relatively good dataset

Tranformers in Action

DALC: Robustness & bias of trained models

- BERTje OFF overgeneralizes on non-hateful cases
- BERTje ABU is more balanced and better on difficult functional tests
- BERTje OFF-ABU: best on non-hateful tests

HATECHECK-NL	
Model	**Accuracy**
Baseline (random)	57.08
BERTje OFF	**61.40**
BERTje ABU	60.19
BERTje OFF-ABU	60.94

We happy?

We happy?
Again, not really

- The pre-training material of BERTje is good quality, but not really suitable for detecting socially unacceptable language
- Solution: adapt the model
  - further pre-training paradigm (Gururangan et al., 2020)
  - unsupervised domain/language variety adaptation
  - early stage adaptation —> cheaper & faster

HateBERT

Evil for good

- A further pre-trained model using instances of “evil communities” from Reddits
  - 29 banned Reddit community (2015)
  - >1M message
  - > 43M tokens
- 100 epochs further pre-training with MLM
  - ~3 weeks - 1 NVIDA V100 GPU
- > 640k downloads and 200 citations

https://ignaco.com/bert/

HateBERT

Banned community overview

Subreddit	Number of posts
apewrangling	5
beatingfaggots	3
blackpeoplehate	16
chicongo	15
chimpmusic	35
didntdonuffins	22
fatpeoplehate	1465531
funnyniggers	29
gibsmedat	24
hitler	297
holocaust	4946
kike	1
klukluxklan	1
milliondollarextreme	9543

Subreddit	Number of posts
misogyny	390
muhdick	15
nazi	1103
niggas	86
niggerhistorymonth	28
niggerrebooted	5
niggerspics	449
niggersstories	75
niggervideos	311
niglets	27
pol	80
polacks	151
sjwhate	10080
teenapers	23
whitesarecriminals	15

HateBERT

Intrinsic evaluation

- Template filling:
  - `[someone] is a(n)/ are [MASK]`
  - `[someone]`= you |she | he | women | men

BERT	HateBERT
"women"	
excluded (.075)	stu**d (.188)
encouraged (.032)	du*b (.128)
included (.027)	id***s (.075)

- HateBERT consistently presents more profanities and slurs as masked terms

HateBERT

Extrinsic evaluation

- 3 datasets in English:
  - **OffensEval 2019** (Zampieri et al., 2019): offensive language
  - **AbusEval** (Caselli et al., 2020): abusive & offensive language
  - **HatEval** (Basile et al., 2019): hate against women and migrants
- All datasets are based on Twitter/X
- Fine-tuning with same hyper-parameters BERT & HateBERT

HateBERT in Action

Experiment results

- in-domain

Dataset	Model	Macro F1	Pos. class - F1
OffensEval 2019	BERT	.803±.006	.715±.009
	HateBERT	**.809±.008**	**.723±.012**
	Best	.829	.752
AbusEval	BERT	.727±.008	.552±.012
	HateBERT	**.765±.006**	**.623±.010**
	Caselli et al. (2020)	.716±.034	.531
HatEval	BERT	.480±.008	.633±.002
	HateBERT	**.516±.007**	**.645±.001**
	Best	.651	.673

- out-of—domain

Train	Model	OffensEval 2019	AbusEval	HatEval
OffensEval 2019	BERT	–	.726	.545
	HateBERT	.–	.750	.547
AbusEval	BERT	.710	–	.611
	HateBERT	.713	–	.624
HatEval	BERT	.572	.590	–
	HateBERT	.543	.555	–

HateBERT

Summary

- HateBERT always better than BERT
  - New sota results on OffensEval & AbusEval
- HateBERT more robust and better portability
  - Portability is better from more generic to more specific phenomena

And how about LLMs?

Not much out yet to be honest

- Zero-shot prompting (Zampieri et al., 2023)

> *Comments containing any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct are offensive comments. This includes insults, threats, and posts containing profane language or swear words. Comments that do not contain offense or profanity are not offensive.*
> *Question: In one word, is the following comment offensive or not? They should just shut up*
> *Answer: The comment is*

https://www.cambridge.org/core/services/aop-cambridge-core/content/view/2605A4C9E45354D36C0B732B49DB8CA3/S1351324923000517a.pdf/offenseval-2023-offensive-language-identification-in-the-age-of-large-language-models.pdf

LLM for Offensive Language

OffensEval 2023

Model	Macro-F1
OffensEval Rank 1	*0.829*
OffensEval Rank 2	*0.815*
OffensEval Rank 3	*0.814*
Flan-T5	0.793
Llama 2	0.715
Falcon	0.648
MPT	0.547
hateBERT	0.507
T0	0.430
fBERT	0.329
RedPajama	0.267

- LLMs lag behind fine-tuned models
- HateBERT is competitive 🥳
  - pre-trained w/o fine-tuning
  - full control of the pre-trained data
- Best LLM is a encoder-decoder
- Llama 2 very competitive

https://www.cambridge.org/core/services/aop-cambridge-core/content/view/2605A4C9E45354D36C0B732B49DB8CA3/S1351324923000517a.pdf/offenseval-2023-offensive-language-identification-in-the-age-of-large-language-models.pdf

Summary

Take-home messages

- Socially unacceptable language is a complex phenomenon
  - numerous dimensions are involved
  - interdisciplinary
- Creating good datasets is not an easy task
  - sparse
  - risk of bias
  - need context of occurrence

Summary

Take-home messages

- Systems for automatic detection:
  - very good on the NOT class
  - improvements needed for the ABU and OFF classes
- Dynamic & functional benchmarking
  - Expose bias in data and system
  - Test robustness and generalization functionalities of systems
  - For hate speech, it requires language & cultural adaptations

Detection is only one side of the story

Counter narratives

- Chung et al., 2020
  - Multilingual (EN, IT, FR) dataset of counter narratives on hate speech
  - Manually created by experts
  - Goal: automatically generate counter narratives

Hate Speech	Counter-Narrative
Every Muslim is a potential terrorist.	Every Muslim is also a potential peacemaker, doctor, philanthropist... What's your point?

https://i.pinimg.com/originals/9f/24/ed/9f24edf4fb0ccc10eb0435db66378b23.jpg

References

DALC

- Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch Abusive Language Corpus. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pages 54–66, Online. Association for Computational Linguistics.
- Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. “Zo Grof !”: A Comprehensive Corpus for Offensive and Abusive Language in Dutch. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 40–56, Seattle, Washington (Hybrid). Association for Computational Linguistics.
- Dion Theodoridis and Tommaso Caselli. 2022. All That Glitters is Not Gold: Transfer-learning for Offensive Language Detection in Dutch. Computational Linguistics in the Netherlands Journal, 12, 141–164
- Tommaso Caselli and Hylke Van Der Veen. 2023. Benchmarking Offensive and Abusive Language in Dutch Tweets. In The 7th Workshop on Online Abuse and Harms (WOAH), pages 69–84, Toronto, Canada. Association for Computational Linguistics.

- DALC - Data: https://doi.org/10.34894/HOINL3
  Code: https://github.com/tommasoc80/DALC/tree/master
- Models: https://huggingface.co/GroNLP

References

AbusEval, HateBERT

- Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I Feel Offended, Don't Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6193–6202, Marseille, France. European Language Resources Association.

- Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for Abusive Language Detection in English. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), pages 17–25, Online. Association for Computational Linguistics.

- Model: https://huggingface.co/GroNLP/hateBERT