Evaluating Quality of Abusive Language Datasets

Chaewon Yun

Seminar: Current Topics in Applied Computational Social Science

Table of Contents

Introduction

Relevant Definitions

What is abusive language dataset?

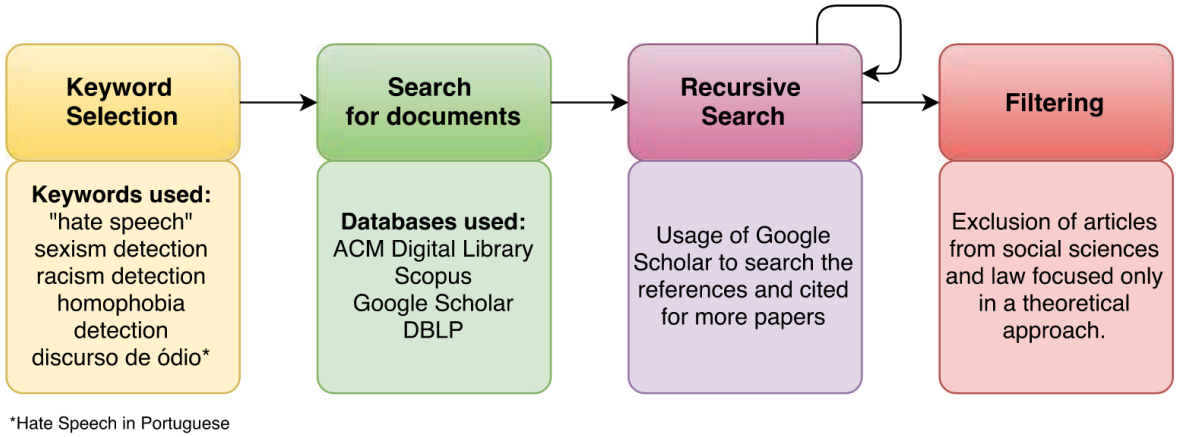


Fig. 1. Methodology for document collection.

Fortuna and Nune (2018)

text (string)	user_id (int)	subforum_id (int)
As of March 13th , 2014 , the booklet had been downloaded over 18,300 times and counting .	572,066	1,346
In order to help increase the booklets downloads , it would be great if all Stormfronters who had…	572,066	1,346
(Simply copy and paste the following text into your YouTube videos description boxes.)	572,066	1,346
Click below for a FREE download of a colorfully illustrated 132 page e-book on the Zionist-…	572,066	1,346
Click on the `` DOWNLOAD (7.42 MB) '' green banner link .	572,066	1,346
Booklet updated on Feb. 14th , 2014 .	572,066	1,346

https://huggingface.co/datasets/hate_speech18

Term: Hate speech, Abusive language, Toxicity..

Fig. 1 Relations between HS and related concepts

Poletto et al. 2021

How to Operationalize the Quality?

Which aspect of quality matters for the hate speech dataset?

- Representativeness
  - What we want: Representative sample of hate speech
  - If not possible, we want at least diversity → avoid bias!
- Fits the purpose of the dataset: Hate speech detection
  - Using supervised learning method
  - Number of instances per label
  - Balanced classes
- Other values
  - Open science practice – e.g., Data availability, Degradation
  - Ethical consideration – e.g., Protect privacy

Input paper

Bias and Comparison Framework for Abusive Language Detection

Maximilian Wich, Tobias Eder, Hala Al Kuwatly, and Georg Groh.

AI and Ethics, 2021

Suggests:
Bias Framework for Abusive Language Datasets

Perspective	Method	Problem
1. Meta	(a) Class distribution and availability	Degradation
	(b) Time distribution	Temporal bias
	(c) Pareto analysis of authors	Author bias
2. Semantic	(a) LSI-based intra-dataset class similarity	Similarity/dissimilarity of classes
	(b) Word embedding based intra- and inter-dataset class similarity	Similarity/dissimilarity of classes
	(c) Cross-dataset topic model	Topic bias
	(d) PMI-Based word ranking for class	Topic bias
3. Annotation	(a) Distribution of inter-rater reliability	Annotator bias
4. Classification	(a) Cross-dataset performance	Generalizability
	(b) Explainable classification models	Generalizability

Wich et al., 2021

Then, apply the framework to these datasets:

Table 2 Selected abusive language datasets (class names in bold are the abusive categories)

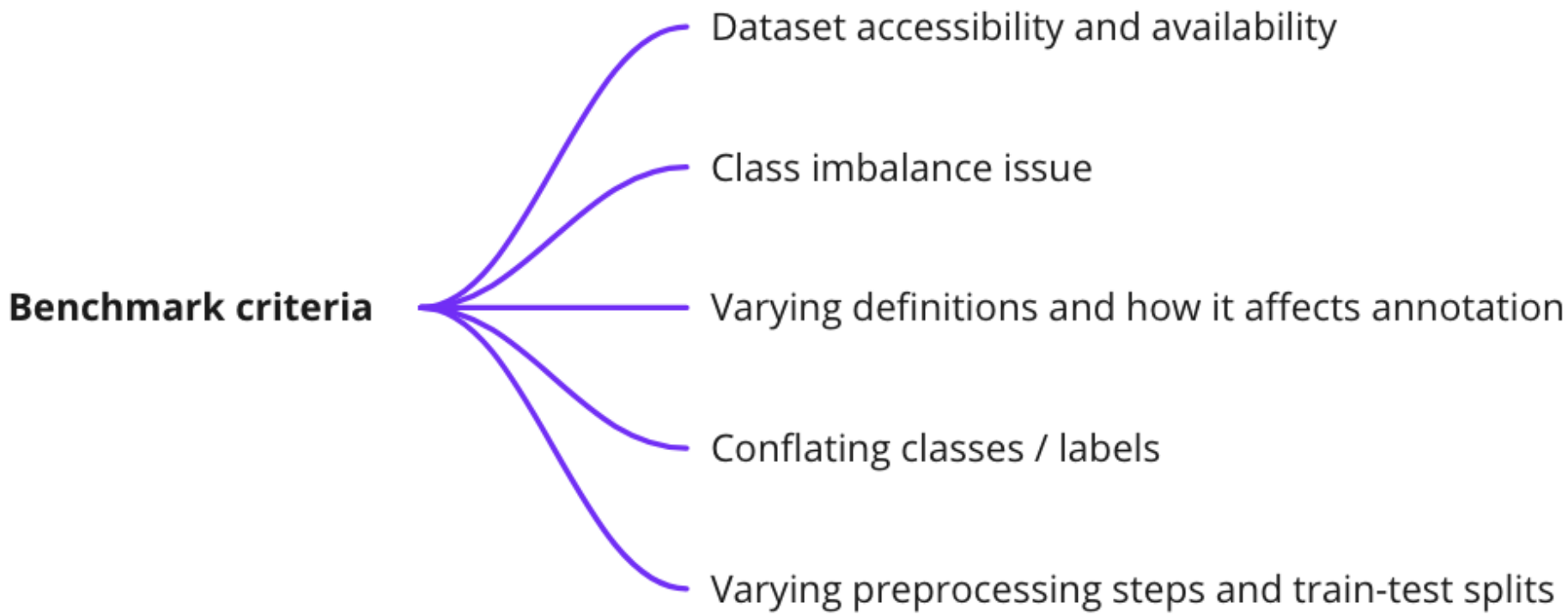
Lang.	Name	Source	Size	Labels	Ref.
English	Waseem	Twitter	16,907	None, sexism, racsim	[44]
	Davidson	Twitter	24,783	Offensive, hate, neither	[10]
	Founta	Twitter	99,996	Normal, abusive, hateful, spam	[16]
	Zampieri	Twitter	14,100	Hierarchical labels: (1) not offensive, offensive (2) if offensive: targeted insult, untargeted insult (3) if targeted: individual target, group target, other	[48]
	Vidgen	Twitter	20,000	Hostility, criticism, counter speech, discussion of East Asian prejudice, neutral	[40]
Arabic	Alsafari	Twitter	5341	3-class: clean, offensive, hate; 6-class: clean, offensive, religious hate, gender hate, nationality hate, ethnicity hate	[3]
	Alshalan	Twitter	8958	Hate, non-hate	[4]
	Albadi	Twitter	6136	Hierarchical labels: (1) neutral, religious hate (2) if religious hate: Muslims, Jews, Christians, Atheists, Sunnis, Shia, other	[2]
	Chowdhury	Twitter, Facebook, YouTube	4000	Hierarchical labels: (1) non-offensive, offensive (2) if offensive: vulgar, hate, only offensive	[7]
	Mubarak	Twitter	9996	Hierarchical labels: (1) non-offensive, offensive (2) if offensive: hate speech, not hate speech	[28]
	Mulki	Twitter	5846	Normal, busive, hate	[29]

Wich et al., 2021

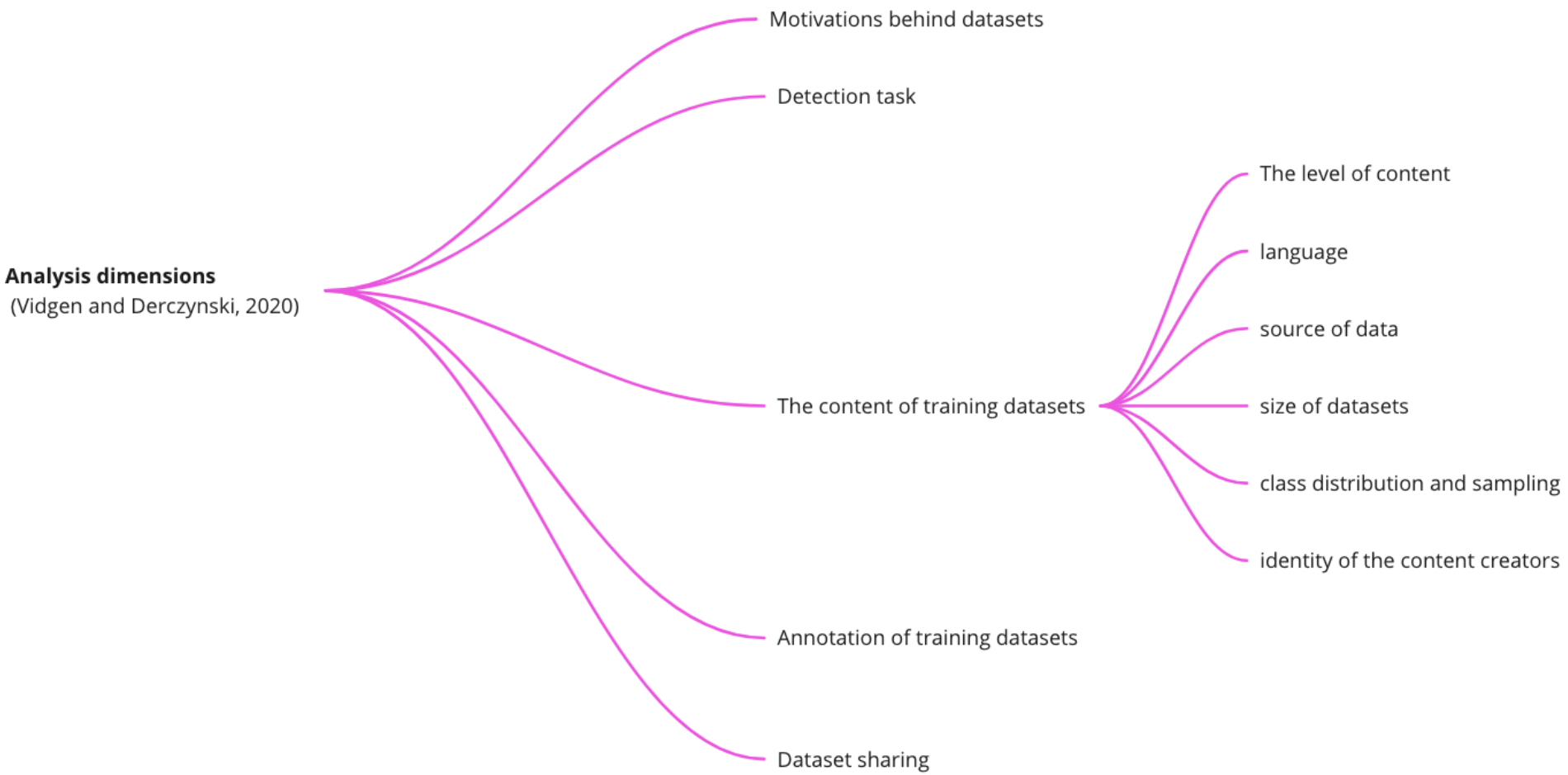
Related Works

Comparing four relevant works on hate speech dataset

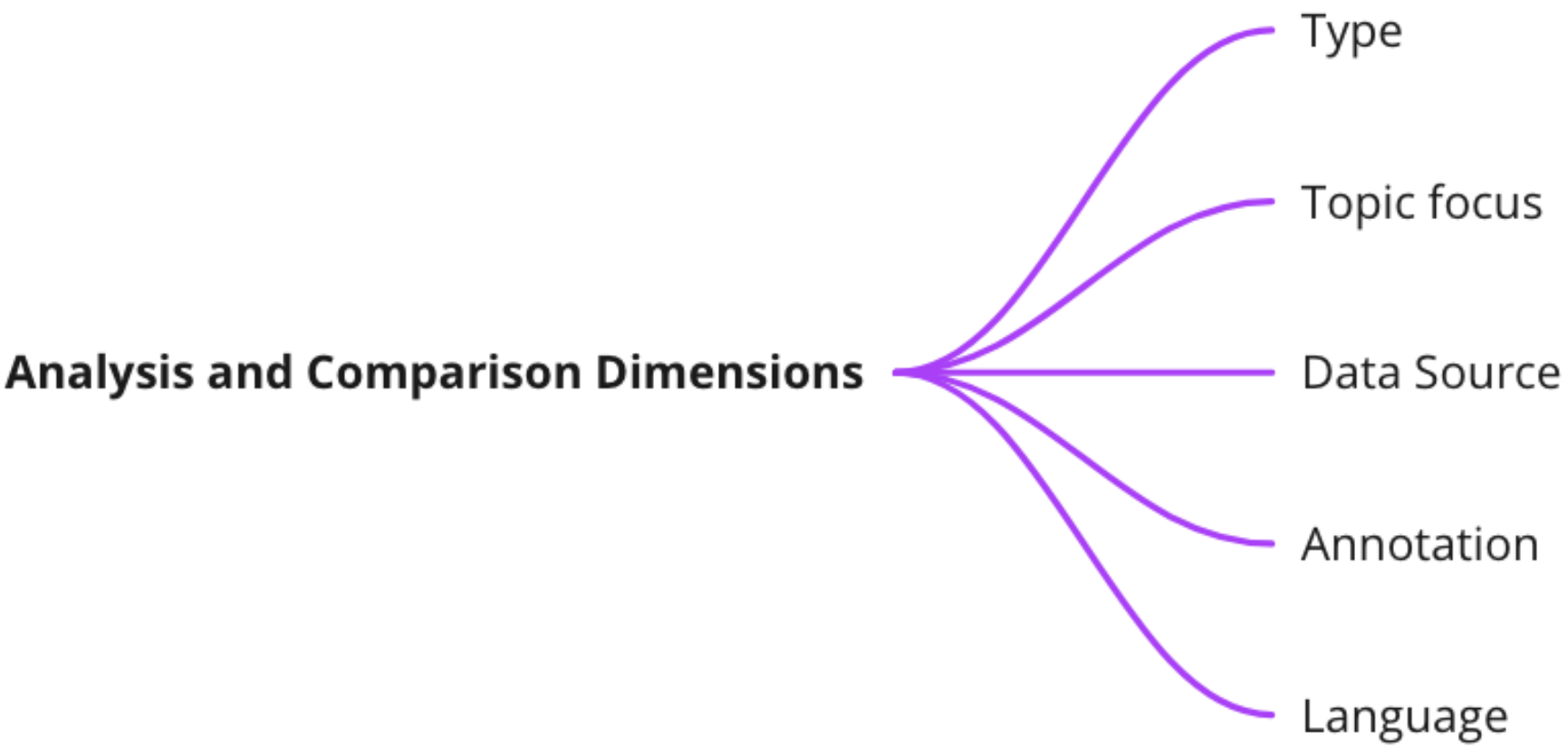
1. “In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets.” (Madukwe et al., 2020)



2. Directions in abusive language training data, a systematic review: Garbage in, garbage out (Vidgen and Derczynski, 2020)



3. Resources and benchmark corpora for hate speech detection: a systematic review. (Poletto et al., 2020)



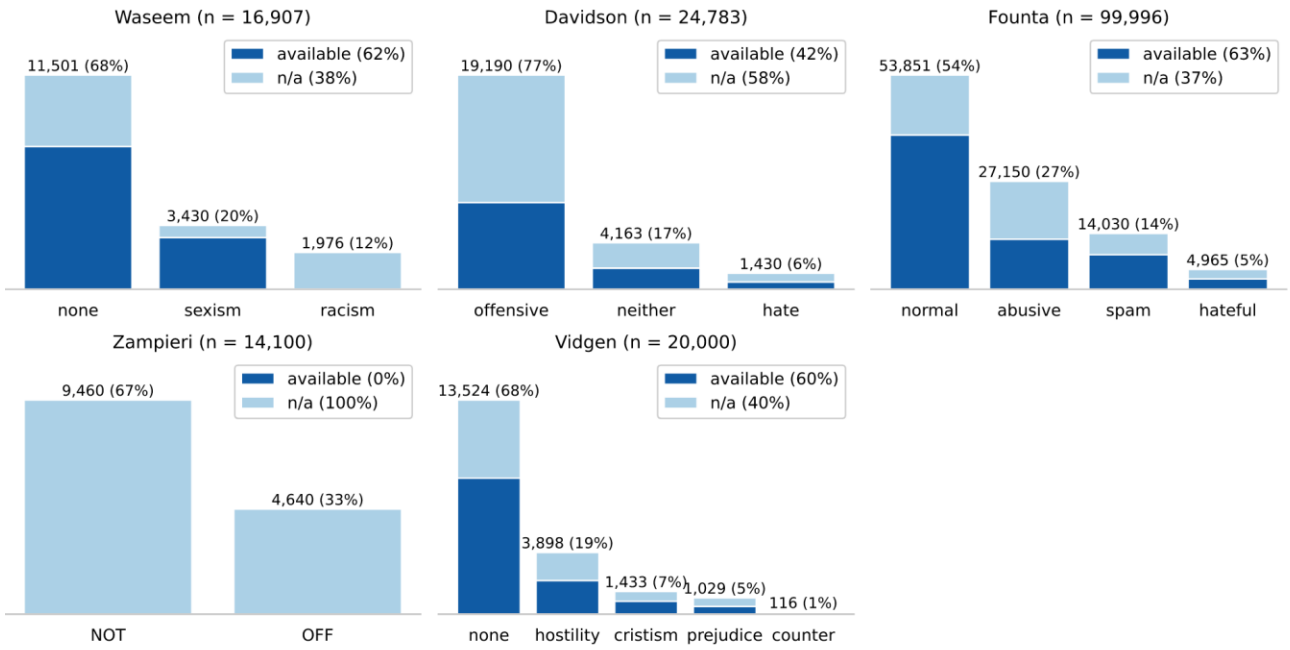
Quality Criteria

How to operationalize dimensions of quality?

Overview of Properties

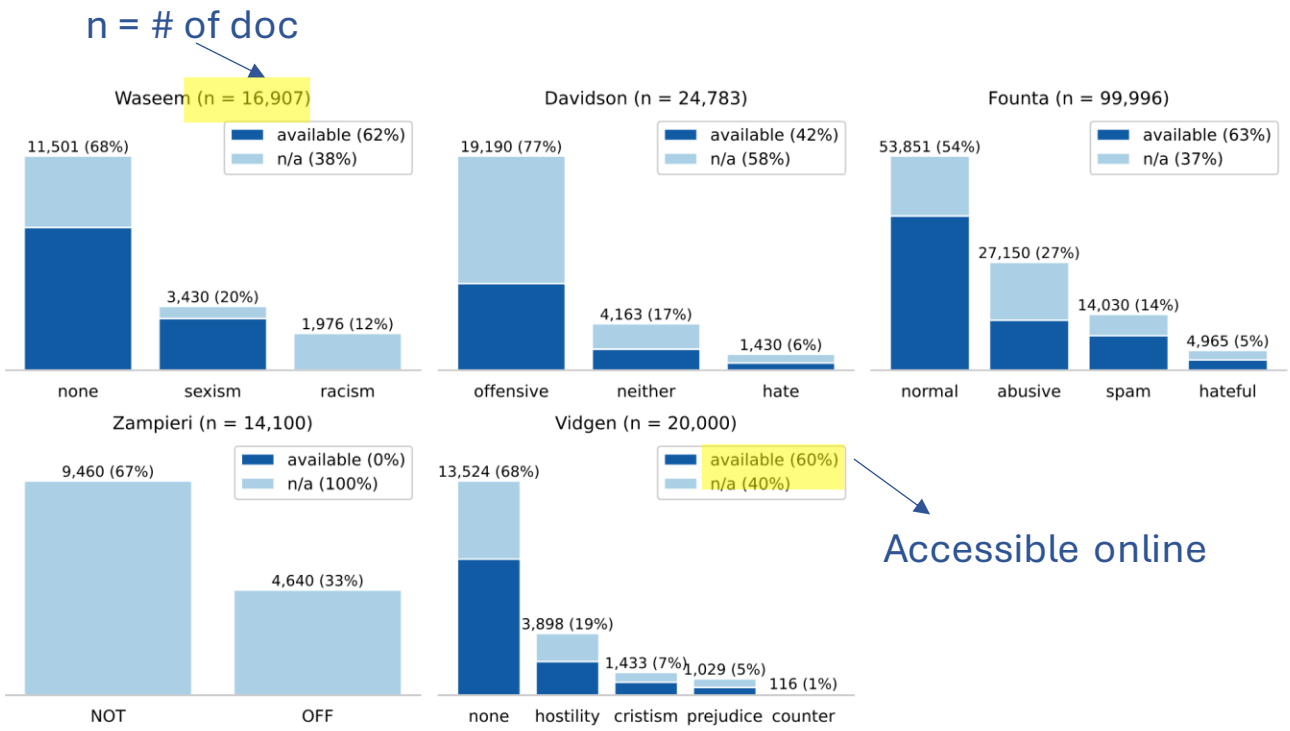
1. Annotation Integrity
   - Inter-Rater / -Annotator Reliability and Consistent Annotation
   - Varying definition across datasets
   - Granularity of Annotation
2. Class Imbalance and Skewed Distribution
3. Bias – Author, Temporal, Topic Bias
4. Generalizability – Cross-dataset performance
5. Data Availability and Degradation

Class distribution (and availability)



Wich et al., 2021

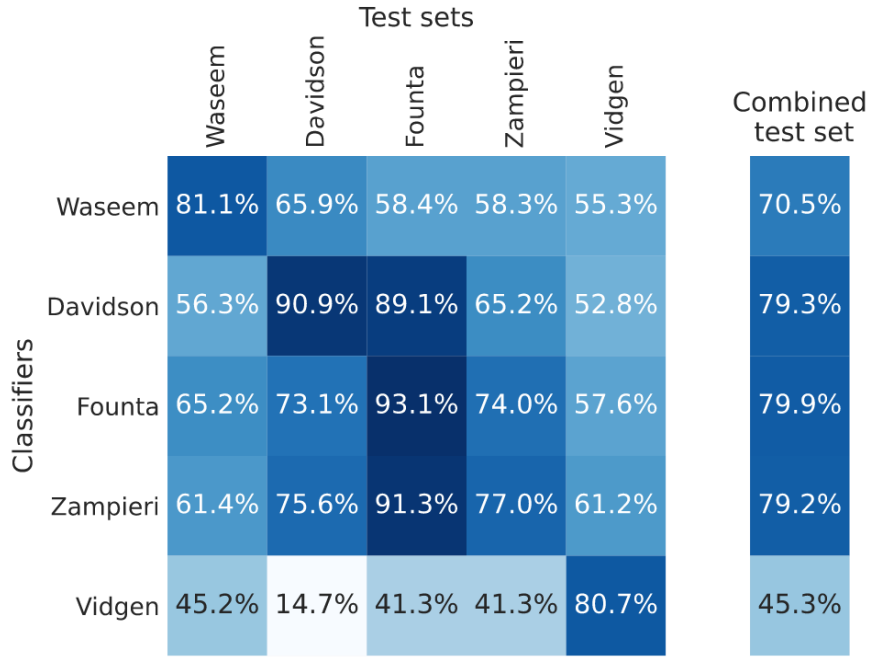
Class distribution (and availability) → Better Training



Wich et al., 2021

Cross-dataset classification performance

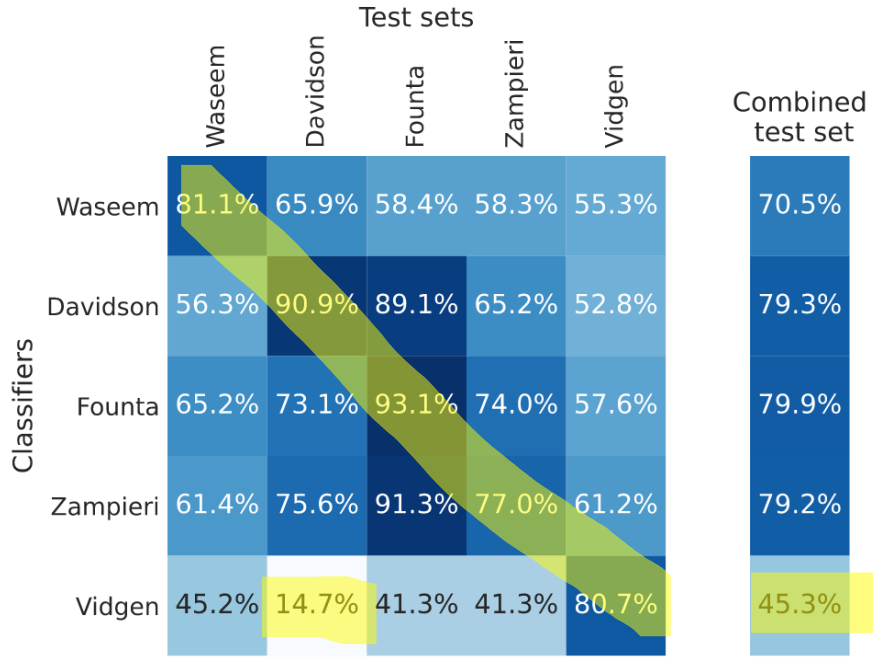
- Trained on different datasets and tested on all test sets.
- As the basis for the classification model, we use the English/Arabic pre-trained BERT
- Vidgen delivers the worst performance, but this should not surprising due to the topic focus.



Wich et al., 2021

Cross-dataset classification performance → Generalizability

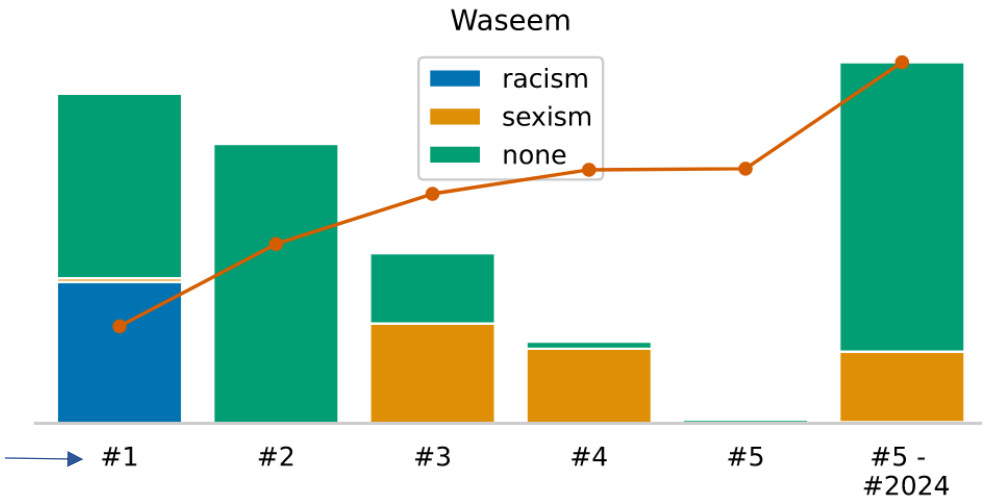
- Trained on different datasets and tested on all test sets.
- As the basis for the classification model, we use the English/Arabic pre-trained BERT
- Vidgen delivers the worst performance, but this should not surprising due to the topic focus.



Wich et al., 2021

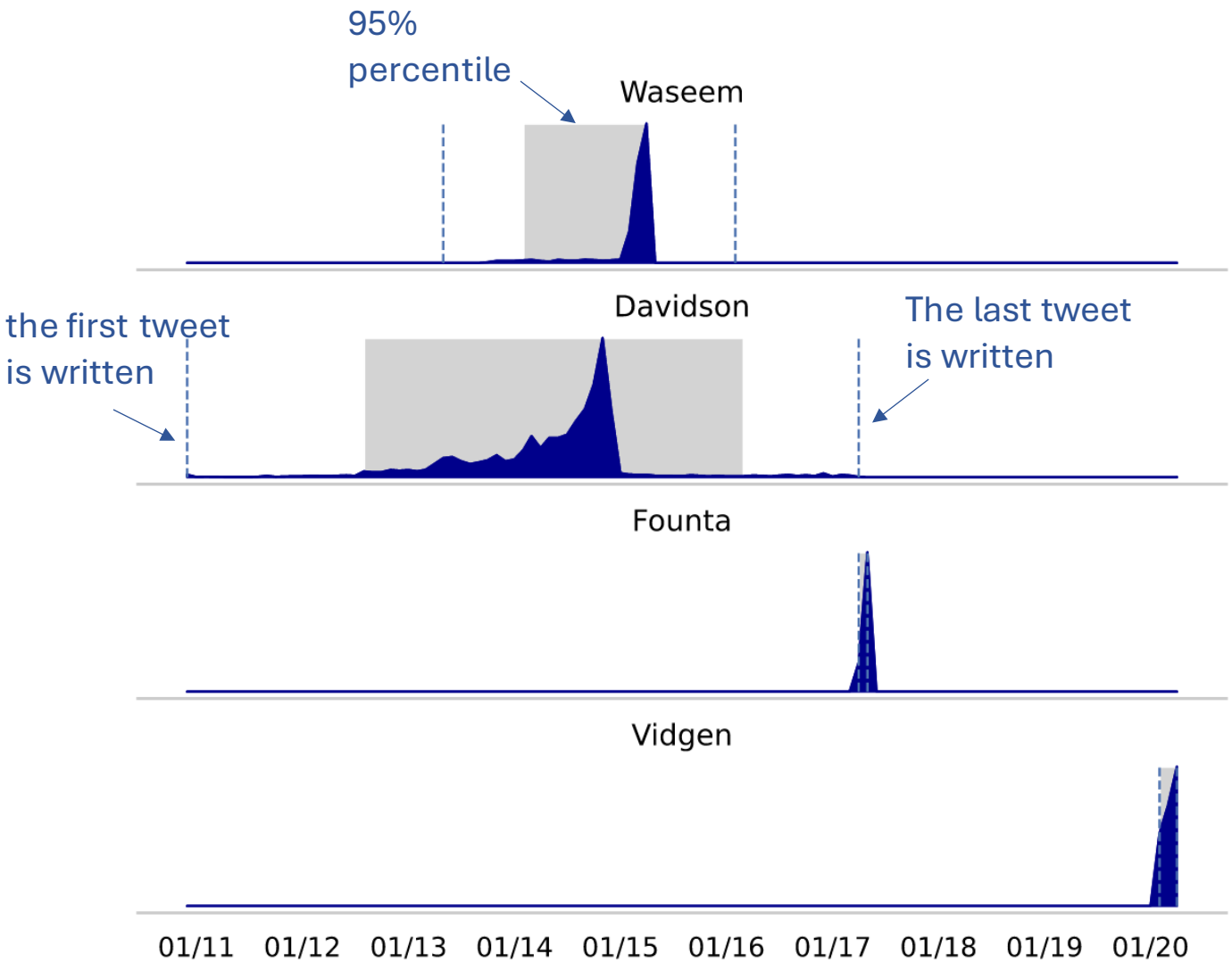
Author Bias → Diversity (Possible confounder)

Top authors of the dataset →



Wich et al., 2021

Temporal bias → Diversity (Possible confounder)



Wich et al., 2021

Discussion

Is it possible to build a benchmark dataset for hate speech detection?

Hate Speech Benchmark?

- Most researchers seem to agree to the necessity of the benchmark dataset
- But is it possible to build one general hate speech dataset to compare various models?
- Special characteristic of hate speech – linguistic nuances, strong cultural/temporal implication, etc.

References

- Paula Fortuna and S´ergio Nunes. A survey on automatic detection of hate speech in text. ACM Computing Surveys (CSUR), 51(4):1–30, 2018.
- Kosisochukwu Judith Madukwe, Xiaoying Gao, and Bing Xue. In Data We Trust: A Critical Analysis of Hate Speech Detection Datasets. In Proceedings of the Fourth Workshop on Online Abuse and Harms, Online: Association for Computational Linguistics, (150–61):1–22, 2020.
- Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, , and Viviana Patti. Resources and Benchmark Corpora for Hate Speech Detection: A Systematic Review. Language Resources and Evaluation, 55(2):477–523., June 2021.
- Anna Schmidt and Michael Wiegand. A Survey on hate speech detection using natural language processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Medias, Association for Computational Linguistics, Valencia, Spain, (150–61):1–10, 2017.
- Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. arXiv, 2017
- Vidgen, Bertie, and Leon Derczynski. “Directions in Abusive Language Training Data, a Systematic Review: Garbage in, Garbage Out.” Edited by Natalia Grabar. *PLOS ONE* 15, no. 12 (December 28, 2020): e0243300. https://doi.org/10.1371/journal.pone.0243300.
- Wich, Maximilian, Tobias Eder, Hala Al Kuwatly, and Georg Groh. “Bias and Comparison Framework for Abusive Language Datasets.” *AI and Ethics*, July 19, 2021. https://doi.org/10.1007/s43681-021-00081-0.

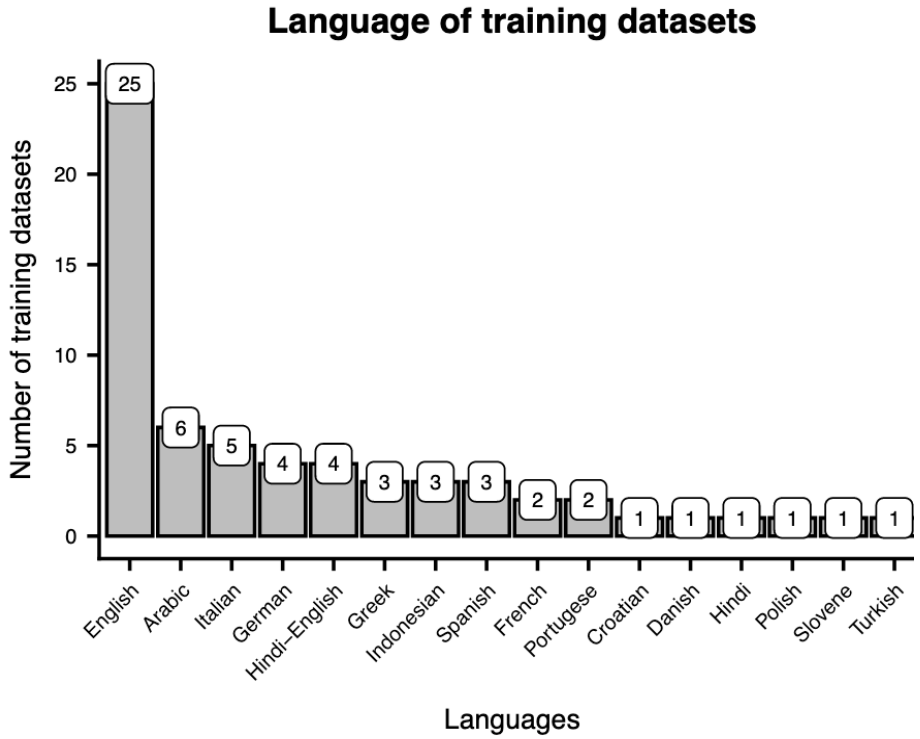
Appendix

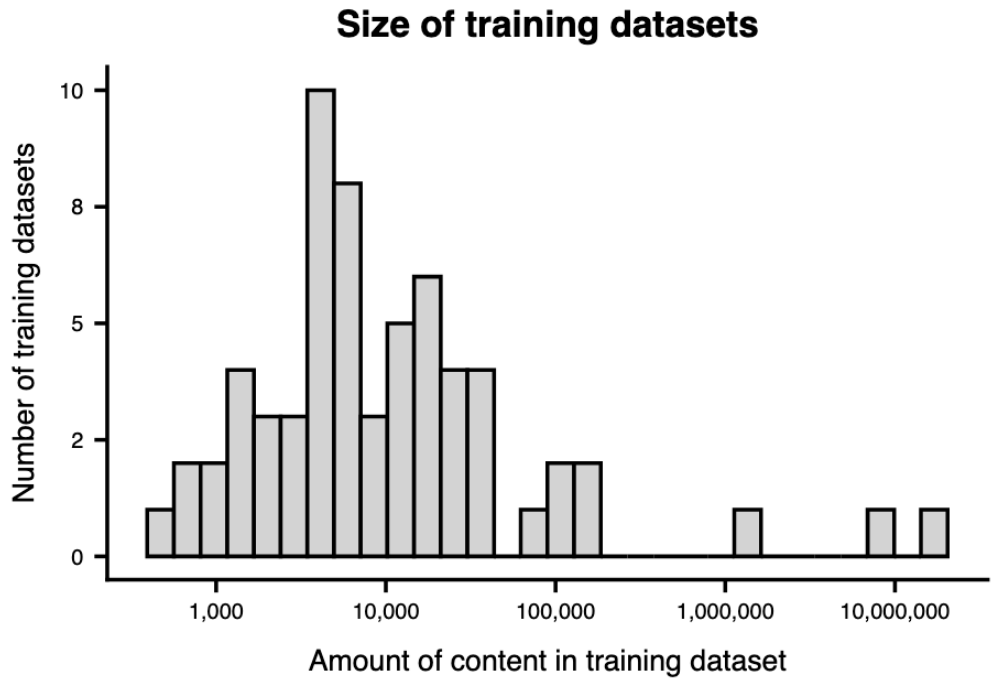
Challenges in Hate Speech Detection

- Inter-rater (Inter-annotator) reliability
- Varying definition
- Varying annotation schema
- Dataset availability
- Cultural, linguistic nuances
- Skewed distribution within the data
- Lack of benchmark dataset

Language and Size of the Datasets

(Vidgen and Derczynski, 2020)

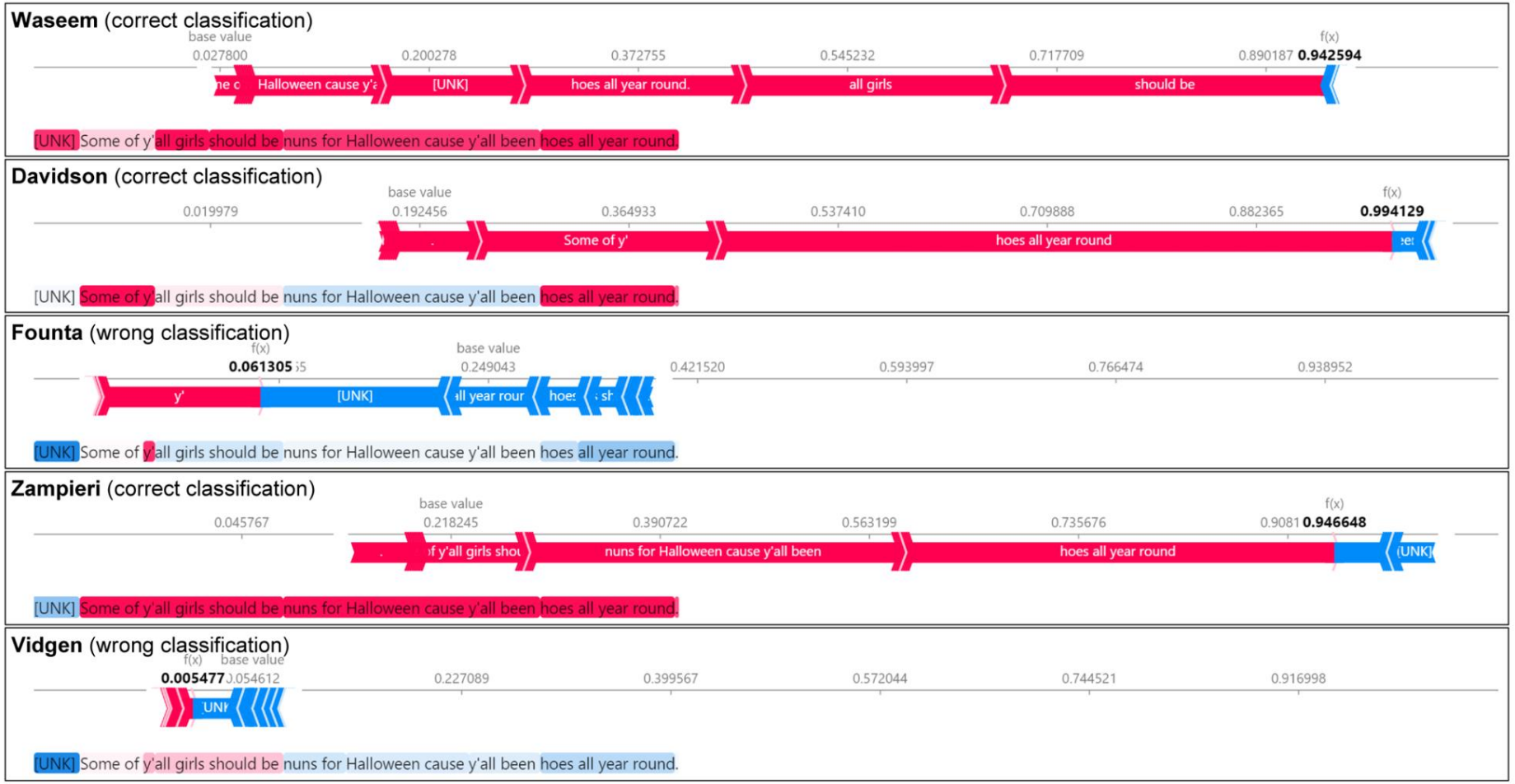




Varying Preprocessing and Train-test split
(Madukwe et al., 2020)

Datasets	Paper	Stem or Lemmatize	Username	URLs	Lowercase	Hashtags	Remove Punctuation	Remove Stopwords	Train-Test Split	Final DataSize
WASEEM	(Badjatiya et al., 2017)	-	replaced	replaced	added <allcaps> after an all capitalized word	replaced # sign with <hashtag>	No. Repetition replaced with <repeat>	-	-	-
	(Founta et al., 2019)	both	counted	counted	No. counted all capital words	counted	-	Yes	-	16,059
	(Mozafari et al., 2020)	-	replaced with placeholder <user>	replaced with placeholder <url>	Yes	removed # sign only	Yes	No	-	-
DAVIDSON	(Davidson et al., 2017)	stem	counted	counted	Yes	counted	-	-	5-fold CV	24,802
	(Malmasi and Zampieri, 2017, 2018)	-	removed	removed	Yes	-	-	-	10-fold CV	14,509
	(Founta et al., 2019)	both	counted	counted	No. counted all capital words	counted	-	Yes	-	24,783
	(Madukwe and Gao, 2019)	lemmatize	removed	removed	Yes	removed	Yes	Yes. Custom stop words	75/25	-
	(Mozafari et al., 2020)	-	replaced with placeholder <user>	replaced with placeholder <url>	Yes	removed # sign only	Yes	No	-	-
	(Miok et al., 2019)	lemmatize	remove	remove	-	expanded into words	Yes	Yes	-	3000
FOUNTA	(Verma et al., 2020)	-	replaced	replaced	Yes	Dropped # sign only	No	Yes	80/10/10	-
	(Liu et al., 2020)	-	-	removed	-	removed	Yes	-	80% 20%	99603 from 100000
	(Davidson et al., 2017)	-	replaced	replaced	-	-	Yes	Yes	-	75,023 from 100000
	(Kim et al., 2020)	Stem	-	-	Yes	-	-	-	80 /20	-

Explainability (Wich et al., 2021)



More discussion questions

- Samples - High diversity vs. Representative?
- Any other overlooked criteria?
- What's remaining open questions?
- What's describable?
- How to avoid confounder?
- Heuristics, ad hoc..

Vidgen and Derczynski, 2020

Anlysis of Training Datasets	Types	
Motivations behind datasets	reducing harm	
	removing illegal content	
	Improving health of online conversations	
	Reducing the burden on human moderators	
detection tasks	The nature of abuse	person-directed, group-directed, flagged, incivil, mixed
	The granularity of taxonomies	binary, multi-level, compex classes
The content of training datasets	The 'level' of content	level of post, user, context information
	Language	
	Source of data	twitter, youtube, world of warcraft, etc.
	Size of datasets	469 post ~ 17 million
	Class distribution and sampling	
	Identity of the content creators	
Annotation of training datasets	Annotation process	experts, crowdsourcing, professional moderators, synthetic data, etc.
	Identity of the annotators	demographic information, expertise and experience, personal experience of abuse, guideline for annotation
Dataset sharing	Data availability	

Madukwe et al., 2020

Datasets	Publicly Available	Consistent Split	Accessible data format	Common Evaluation Metric	Unbiased	Pre-processed
WASEEM	✓	✗	✗	✗	✗	✗
DAVIDSON	✓	✗	✓	✗	✗	✗
FOUNTA	✓	✗	✗	✗	✗	✗
QIAN	✓	✗	✓	✗	✗	✗
HATEVAL	✓	✗	✓	✗	✗	✗

Table 3: Benchmark criteria met by datasets

Poletto et al., 2020

- TYPE: what is the structure of the resource;
- TOPICAL FOCUS: how HS and related phenomena are distinguished according to their topical focus or targets, and to what extent such topics or targets are studied;
- DATA SOURCE: where data have been collected from;
- ANNOTATION: how and by whom data have been labeled, according to what framework, and how quality has been assessed;
- LANGUAGE: how different languages are covered, and how resources and definitions vary across languages.

Name/Reference	Focus	Language	Size	Av.	Cit.
AraHate-BNS/CHI/PMI (Albadi et al. 2018)	HS	ara	1523	Yes	< 50
(Davidson et al. 2017)	HS, racism, sexism, homophobia	eng	179	Yes	< 500
HurtLex (Bassignana et al. 2018)	abusiveness, offensiveness	53 languages	< 100,000	Yes	< 50
(Mubarak et al. 2017)	obscenity, profanity, offensiveness	ara	288	No	< 100
(Olteanu et al. 2018)	HS	eng	163	No	< 50
PeaceTechLab lexicon (Ferroggiaro et al. 2018)	HS	multilingual	< 1000	Yes	n.a.
(Qian et al. 2019b)	HS	eng	2105	No	< 10
(Wiegand et al. 2018a)	abusiveness	eng	1651/8479	Yes	< 50