Equally Safe Online?

46%

of women have received abuse

50%

of Black and minority ethnic women

(Glitch & EVAW, 2020)

23%

of women comfortable expressing political views online

(Stephens et al., 2024)

Automatic Identification and Classification of Misogynistic Language on Twitter

Maria Anzovino[1], Elisabetta Fersini[1(✉)], and Paolo R... [2]

Overview of EXIST 2023: sEXism Identification in Social NeTworks

Laura Plaza[1,2(✉)], Jorge Carrillo-de-Albornoz[1,2], Roser Morante[1], Enrique Amigó[1], Julio Gonzalo[1], Damiano Spina[2], and Paolo Rosso[3]

SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter

Valerio Basile[♢] Cristina Bosco[♢] Elisabetta Fersini[♡]
Debora Nozza[♡] Viviana Patti[♢] Francisco Rangel[♠♣]
Paolo Rosso[♣] Manuela Sanguinetti[♢]

Developing a Multilingual Annotated Corpus of Misogyny and Aggression

Shiladitya Bhattacharya[1], Siddharth Singh[2], Ritesh Kumar[2], Akanksha Bansal[3], Akash Bhagat[2], Yogesh Dawer[2], Bornini Lahiri[4], Atul Kr. Ojha[3]

"Be nice to your wife! The restaurants are closed": Can Gender Stereotype Detection Improve Sexism Classification?

Patricia Chiril and **Farah Benamara** and **Véronique Moriceau**

SWSR: A Chinese dataset and lexicon for online sexism detection

Aiqi Jiang [a,*], Xiaohan Yang [b], Yang Liu [a], Arkaitz Zubiaga [a]

“Computer scientists should stop thinking about online hate speech as something requiring **methods**, and start thinking about it as something that demands **solutions**. This change — treating hate speech less like a task and more like the **real-world problem** it is — would **orient CS research towards the concerns of other stakeholders**, and thus begin the **collaborative pursuit** toward a safe Internet.”

Parker & Ruths, 2020

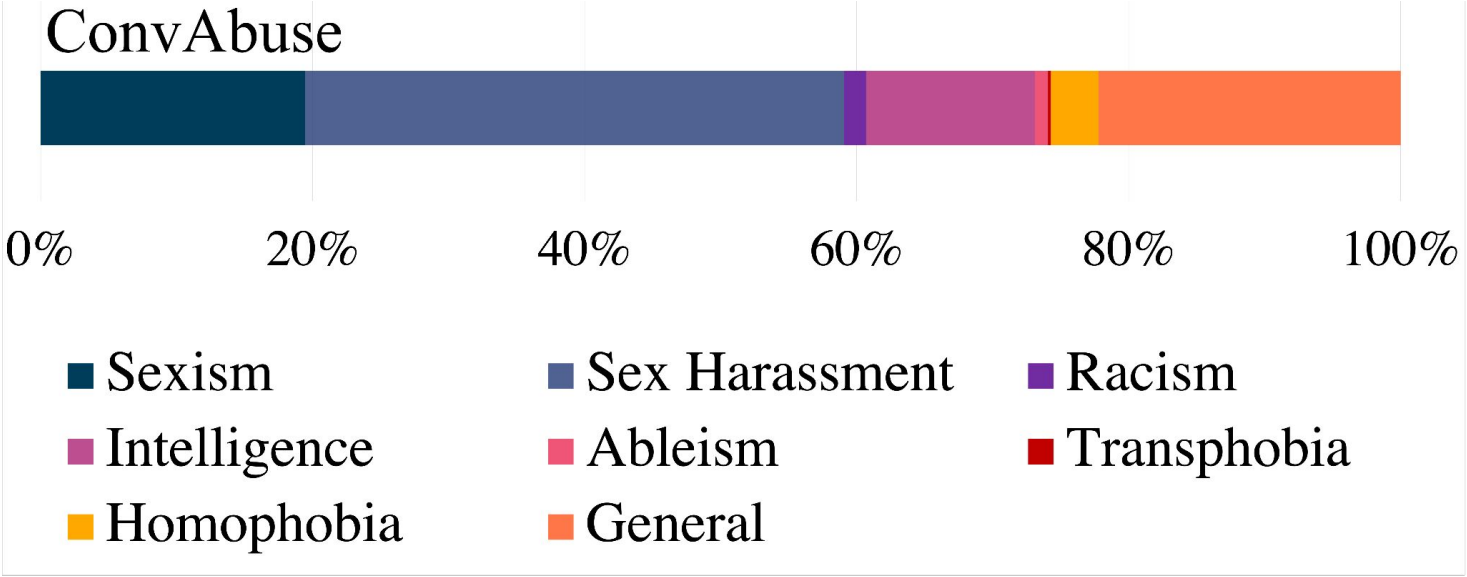
Equally Safe Online

Gavin Abercrombie Heriot-Watt University

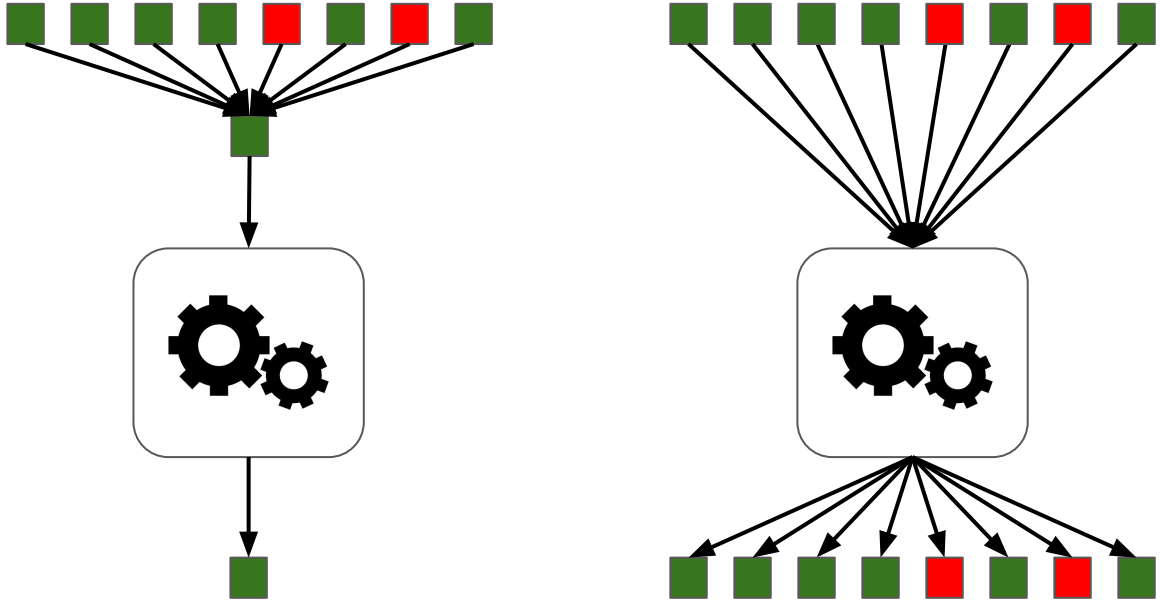
ConvAbuse: Data, Analysis, and Benchmarks for Nuanced Abuse Detection in Conversational AI

Amanda Cercas Curry[1] and **Gavin Abercrombie**[1] and **Verena Rieser**[1,2]

EMNLP 2021


ConvAbuse
0%
20%
40%
60%
80%
100%
Sexism
Sex Harassment
Racism
Intelligence
Ableism
Transphobia
Homophobia
General

$\alpha = 0.63$



‘majority vote aggregation erases minoritized perspectives by choosing as the “right” label whatever is chosen by the majority of annotators, implicitly foregrounding dominant understandings and language ideologies.’

Blodgett, 2021

Say **Anything** with Boyfriend :)

Pron
Adv
Noun

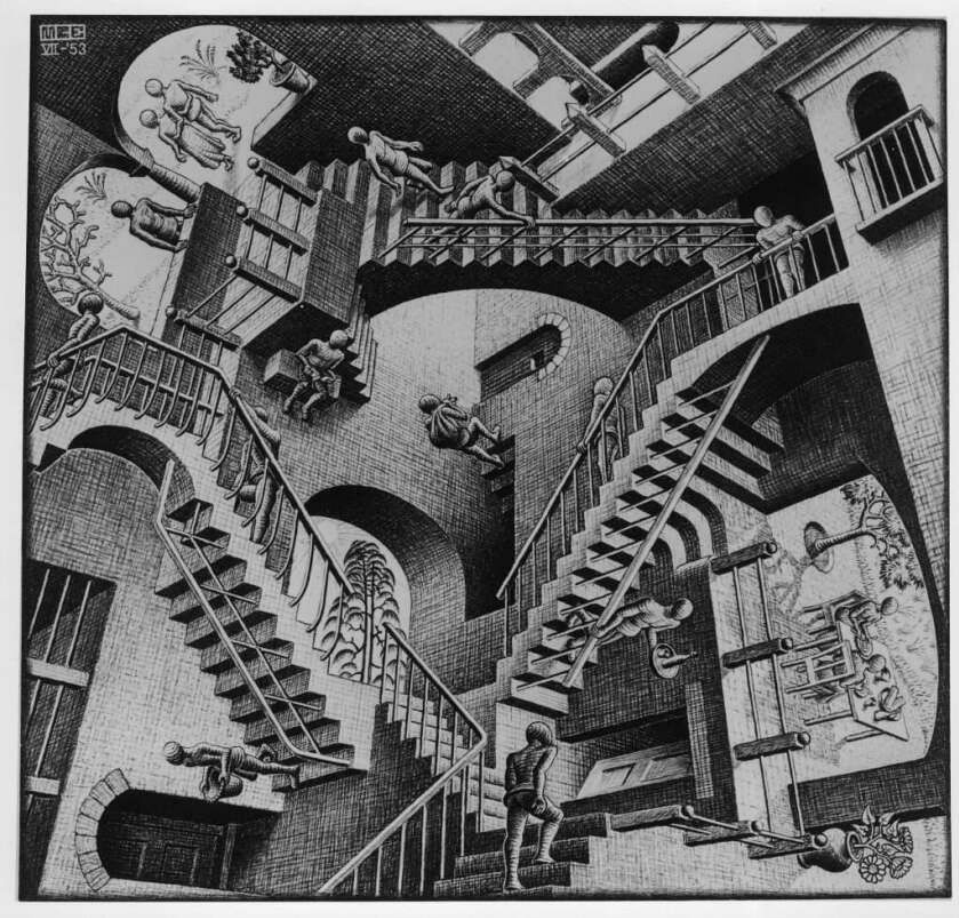
Gimpel et al. 2011.

What is the background metal structure?

Ms COCO image id 393274, VQA 2.0 question id 393274004

1) trees
2) station
3) awning
4) platform
5) platform
6) platform
7) roof
8) shelter
9) train stop
10) awning

Goyal et al. 2017.



https://nlperspectives.di.unito.it/

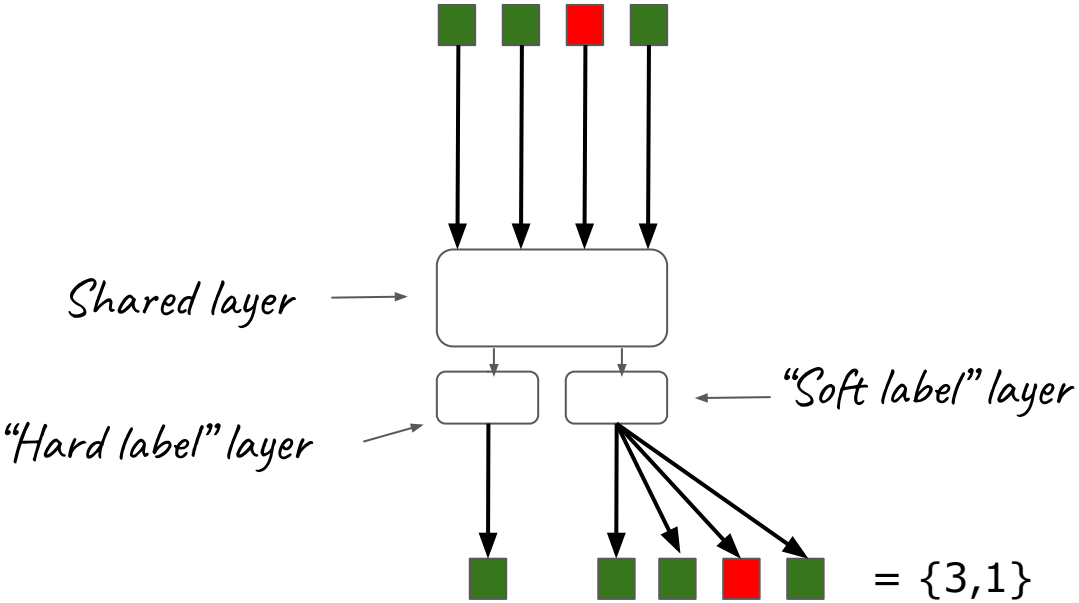
LREC 2022 Marseille

ECAI 2023 Krakow

LREC-COLING 2024 Torino

SemEval-2023 Task 11: Learning With Disagreements (LeWiDi)

Elisa Leonardelli[1] **Gavin Abercrombie**[2] **Dina Almanea**[3]
Valerio Basile [4] **Tommaso Fornaciari**[5] **Barbara Plank** [6]
Verena Rieser[2] **Alexandra Uma** **Massimo Poesio**[3]



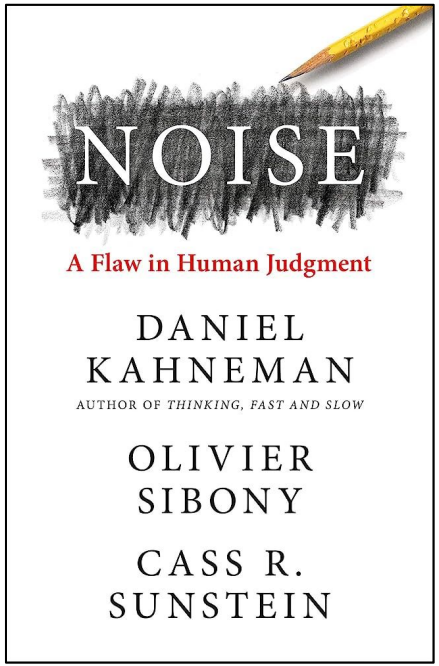
Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling

Gavin Abercrombie
Heriot-Watt University

Dirk Hovy
Bocconi University

Vinodkumar Prabhakaran
Google Research

LAW 2023



Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling

Gavin Abercrombie
Heriot-Watt University

Dirk Hovy
Bocconi University

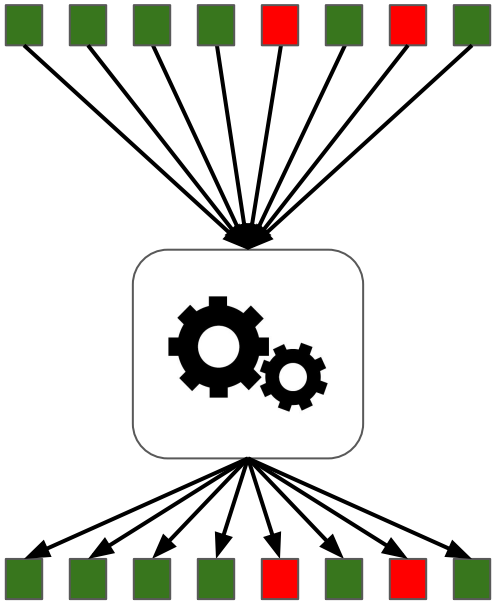
Vinodkumar Prabhakaran
Google Research

74.5%
stability

		Stability (temporal within annotator)	
		High *intra*	Low *intra*
Reliability (between annotators)	High *inter*	Straight-forward / Good quality	Systematic errors/ Value changes
	Low *inter*	Variable perspectives (high subjectivity)	Ambiguous or difficult / Poor quality

Subjective *Isms*? On the Danger of Conflating Hate and Offence in Abusive Language Detection

Amanda Cercas Curry* **Gavin Abercrombie*** **Zeerak Talat***



"When categories of abuse are described as subjective, we understand that there is no ground truth, and wider cultural norms do not impact what constitutes hate. Within the concept of *isms*, we argue that is the wrong approach and that these are culturally defined. That is, we argue that, for a stereotype or norm, there is a ground truth given by the cultural and temporal context a statement is made in."

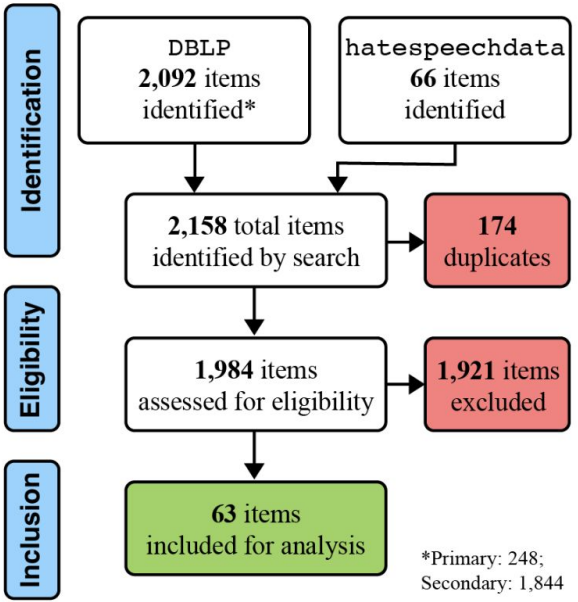
Task formulation
Annotator recruitment

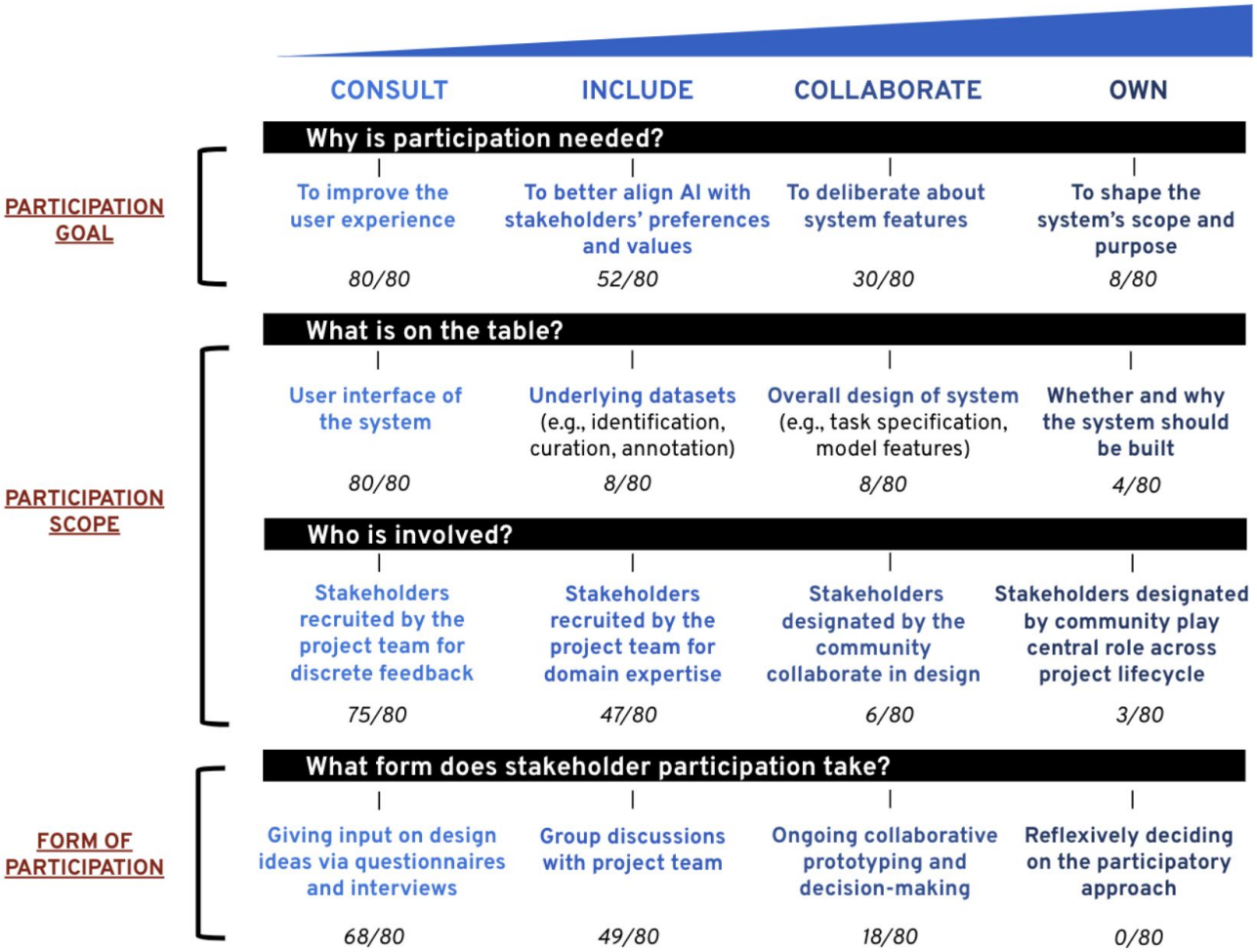
Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review

Gavin Abercrombie[1] and **Aiqi Jiang**[1,3] and **Poppy Gerrard-Abbott**[4,5] and **Ioannis Konstas**[1,2] and **Verena Rieser**[1*]

R1. How is GBV characterised?
R2. Who is represented in annotation of the data?
R3. From which platforms have the data been sourced?
R4. How has the data been sampled?
R5. Which languages are represented?
R6. During which time periods were the data created?

https://github.com/HWU-NLP/GBV-Resources





Delgado et al., 2023

Co-designing A Taxonomy of Online GBV

3 * workshops

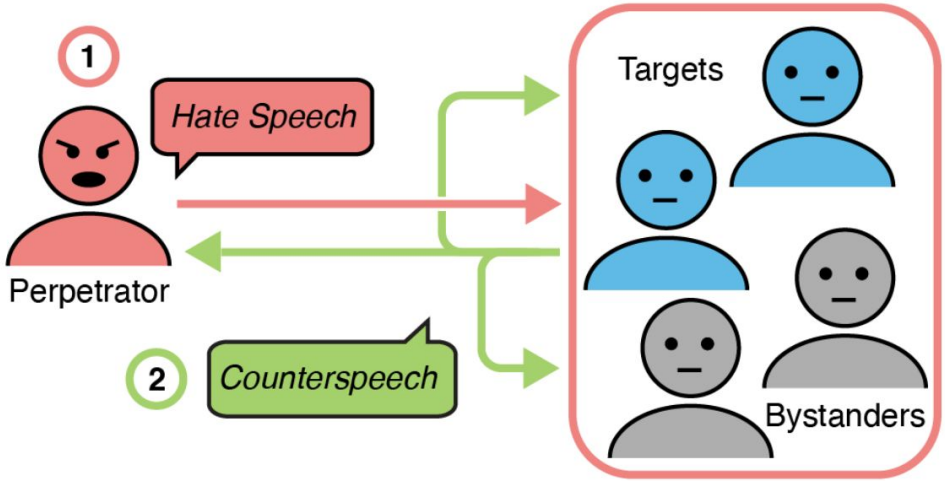
5 * focus groups

Severity cannot be judged from outside

Continuum of disempowerment to empowerment

Understanding Counterspeech for Online Harm Mitigation

Yi-Ling Chung[1] Gavin Abercrombie[2] Florence Enock[1]
Jonathan Bright[1] Verena Rieser[2]*



NLP for Counterspeech against Hate:
A Survey and *How-To* Guide 🔧

Helena Bonaldi[1,2] **Yi-Ling Chung**[3] **Gavin Abercrombie**[4] **Marco Guerini**[1]

A Strategy Labelled Dataset of Counterspeech

Aashima Poudhar[1] and **Ioannis Konstas**[1,2] and **Gavin Abercrombie**[1]

Listen to stakeholders!

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023. Resources for Automated Identification of Online Gender-Based Violence: A Systematic Review. In *WOAH*. Association for Computational Linguistics.

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and Second Language Influence on Intra-Annotator Agreement and Stability in Hate Speech Labelling. In *Proceedings of LAW-XVII*. Association for Computational Linguistics.

Su Lin Blodgett. 2021. *Sociolinguistically Driven Approaches for Just Natural Language Processing.*

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for Counterspeech against Hate: A Survey and How-To Guide. In *Findings of NAACL*. ACL.

Amanda Cercas Curry, Gavin Abercrombie & Verena Rieser. 2021. Convabuse: Data, Analysis & Benchmarks for Nuanced Abuse Detection in Conversational AI. In Proceedings of EMNLP. ACL.

Amanda Cercas Curry, Gavin Abercrombie, and Zeerak Talat. 2024. Subjective Isms? On the Danger of Conflating Hate and Offence in Abusive Language Detection. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 275–282, Mexico City, Mexico. Association for Computational Linguistics.

Chung, Y. L., Abercrombie, G., Enock, F., Bright, J., & Rieser, V. (2023). Understanding counterspeech for online harm mitigation

Delgado, F., Yang, S., Madaio, M., & Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of EEAMO*. ACM.

Gimpel et al. 2011. Part of Speech Tagging for Twitter: Annotation, features, and experiments.

Glitch UK and EVAW. 2020. The ripple effect: COVID-19 and the epidemic of online abuse.

Goyal et al. 2017. *Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering.*

Daniel Kahneman, Olivier Sibony & Cass R. Sunnstein. 2021. *Noise. A Flaw in Human Judgement*.

Elisa Leonardelli, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. ACL.

Aashima Poudhar, Ioannis Konstas, and Gavin Abercrombie. 2024. A Strategy Labelled Dataset of Counterspeech. In *Proceedings of WOAH 2024)*. ACL.

Stevens, F., Enock, F. E., Sippy, T., Bright, J., Cross, M., Johansson, P., ... & Margetts, H. Z. (2024). Women are less comfortable expressing opinions online than men and report heightened fears for safety: Surveying gender differences in experiences of online harms.