Can NLP Tackle Hate Speech in the Real World? Stakeholder-Informed Feedback and Survey on Counterspeech

Tanvi Dinkar[⋆], **Aiqi Jiang**[⋆], **Simona Frenda**[⋆], **Poppy Gerrard-Abbott**[◇], **Nancie A. Gunson**[⋆], **Gavin Abercrombie**[⋆], **Ioannis Konstas**[⋆]
[⋆]Heriot-Watt University, [◇]University of Edinburgh
{T.Dinkar, A.JIANG, S.Frenda, N.Gunson, G.Abercrombie, I.Konstas}@hw.ac.uk
pe.gerrard-abbott@ed.ac.uk

Abstract

Counterspeech, i.e. the practice of responding to online hate speech, has gained traction in NLP as a promising intervention. While early work emphasised collaboration with non-governmental organisation stakeholders, recent research trends have shifted toward automated pipelines that reuse a small set of legacy datasets, often without input from affected communities. This paper presents a systematic review of 74 NLP studies on counterspeech, analysing the extent to which stakeholder participation influences dataset creation, model development, and evaluation. To complement this analysis, we conducted a participatory case study that spanned close to two years with five NGOs specialising in online Gender-Based Violence (oGBV), identifying stakeholder-informed practices for counterspeech generation. Our findings reveal a growing disconnect between current NLP research and the needs of communities most impacted by toxic online content. We conclude with concrete recommendations for re-centring stakeholder expertise in counterspeech research.

Keywords: Counterspeech, Hate Speech, Participatory Design, Systematic Survey

1. Introduction

The automation of counterspeech responses to toxic online content such as hate speech and disinformation is a growing topic in Natural Language Processing (NLP) (Bonaldi et al., 2024a). At the same time, there has been increasing recognition that NLP research should aim to focus on the needs of stakeholders and that the tools it develops should be designed to serve communities (i.e. through participatory design) (Birhane et al., 2022; Caselli et al., 2021), particularly when it comes to tackling hate speech (Abercrombie et al., 2023b; Parker and Ruths, 2023).

Inspired by the work of non-governmental organisations (NGOs) engaged in toxicity countering[1], efforts at automating counterspeech generation began quite promisingly in this regard, with a focus on integrating experts at combating real-world online toxicity into human-in-the-loop systems in the CONAN[2] family of datasets (Bonaldi et al., 2022; Chung et al., 2019; Fanton et al., 2021a). However, as we show in this review, recent work has relied on automated research pipelines in which a few, now relatively old counterspeech datasets are repeatedly reworked with further layers of automatic and/or non-expert produced data, and stakeholders (outwith the computer scientists conducting the research) are typically not involved in their conception, development, or evaluation.

Where recent reviews of counterspeech research have focused on either synthesising findings from real-world counterspeech campaigns (Chung et al., 2024) or technical aspects of natural language generation (Bonaldi et al., 2024a), we focus on stakeholder participation in NLP research in this work.

Our contributions We conduct a **systematic review** (§3) of 74 relevant publications focused on data resources, models, and computational analysis of counterspeech, and answer (**RQ1**): To what extent are affected stakeholders represented in NLP counterspeech research?

We then assess the reviewed work against insights from stakeholders and experts on the best approaches to counterspeech. As a **case study** (§4), we discuss findings from participatory design work spanning nearly two years with five NGOs that work to tackle online Gender-Based Violence (oGBV) in relation to our survey, and investigate (**RQ2**): What stakeholder-informed feedback practices can be used to counter hate?

Findings suggest that NLP research on counterspeech should be redirected towards the needs of such stakeholders. Based on the feedback and issues raised, we provide specific recommendations for NLP practitioners to produce stakeholder-informed counterspeech (§5)[3].

[1]e.g. Get the Trolls Out https://getthetrollsout.org

[2]https://github.com/marcoguerini/CONAN

[3]A full record of the surveyed resources can be found at this link https://github.com/HWU-NLP/CounterspeechResources.git.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8276–8290
11–16 May 2026.

2. Background and Key Concepts

As an alternative to content removal, **Counter-speech** refers to responses that challenge toxic online content, and is seen as a promising way of tackling hate. In NLP, research has focused on creating datasets (Mathew et al., 2018b; Chung et al., 2021c), developing automated counterspeech generation systems (Bonaldi et al., 2023; Gupta et al., 2023), and designing (usually intrinsic) evaluation methods (Zubiaga et al., 2024a; Halim et al., 2023). In sociology, Buerger and Wright (2019) and Alsagheer et al. (2022) review recent trends in counterspeech and provide general introductions to its concept, features and applications, while Benesch et al. (2016) propose a taxonomy of strategies used to counter hate online. From an NLP perspective, Chung et al. (2024) survey the dynamics and effectiveness of counterspeech, and Bonaldi et al. (2024a) the methods and challenges involved in its automation. Tomalin and Ullmann (2023) contribute by compiling multidisciplinary perspectives on counterspeech, including its automation and evaluation. This survey addresses existing gaps by highlighting the importance of stakeholder perspectives in developing counterspeech.

The growing application of AI systems for social good (Moorosi et al., 2023) has increased the engagement of stakeholders in research; with different structures, principles and modalities to guide **participatory design** (Caselli et al., 2021; Birhane et al., 2022; Delgado et al., 2023). However, Parker and Ruths (2023) have identified a disconnect between computer science research and affected communities when it comes to tackling hate speech and its consequences. They propose key points to create a more integrated community to address this: involving groups that combat hate speech who have a deeper understanding of responses to hate speech and its impact on society. In this context, participatory design, popular in branches of computer science such as human-computer interaction (Muller and Kuhn, 1993), gives a voice in the design process to people who lack expert design skills.

Whilst not explicitly referencing participatory methodologies, several early NLP works on counterspeech engaged with domain expert stakeholders to create human-in-the-loop generation pipelines (Chung et al., 2019; Bonaldi et al., 2022; Fanton et al., 2021b). More recently, Mun et al. (2024a) conducted a large-scale survey with relevant stakeholders to inform the design of NLP counterspeech tools. In this work, we uncover the extent to which stakeholders participate in NLP counterspeech research design and resource creation.

Online Gender-Based Violence or *oGBV* is a framework used by international organisations such as the UN and WHO, and covers harmful effects on all genders, particularly women.[4] Misogynistic abuse affects around 50% of women and especially further marginalised groups (Glitch, 2020; Parikh et al., 2019), resulting in women often feeling uncomfortable online (Stevens et al., 2024). Although there have been recent efforts to identify oGBV, including various SEMEVAL tasks (Basile et al., 2019; Fersini et al., 2022; Kirk et al., 2023), existing computational approaches and datasets suffer from several shortcomings (Abercrombie et al., 2023b), such as the lack of participation in designing taxonomies and formalisms of the addressed social problem, and the exclusion, due to the adopted terminology, of specific aspects related to various forms of violence. Our case study describes feedback of stakeholders in addressing oGBV, from focus groups that involved survivors and professional supporters of victims.

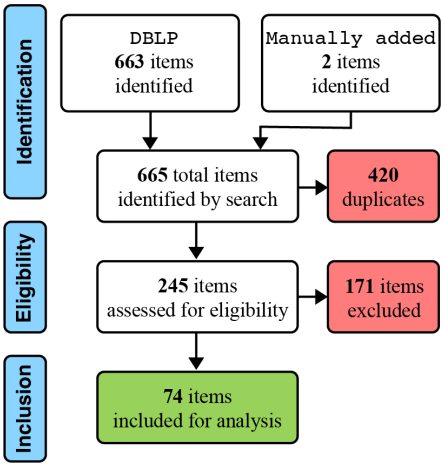

Figure 1: Search and selection protocol.

3. Systematic Review

We conducted a systematic review of computer science publications on the topic of counterspeech, following the PRISMA methodology (Moher et al., 2009). The review protocol is shown in Figure 1.

Identification To isolate relevant counterspeech research and exclude work from fields such as social science that are not concerned with NLP methods, we searched the computer science bibliography database DBLP. All searches were conducted in March 2025. Following Chung et al. (2024), we used the keywords '*counter-speech*', '*counter-narratives*', '*counter-terrorism*', '*counter-aggression*', '*counter-hate*', '*counter speech*', '*counter narrative*', '*countering online hate speech*', '*counter hate speech*', and '*counter-hate speech*', and additionally added the keyword '*counterspeech*'.

[4] https://www.who.int/health-topics/violence-against-women

Publication	HS source	CS source	Human input and Task (None = ×)	Stakeholder involvement (✓/×) with Details
♡ CONAN (Chung et al., 2019)	Nichesourcing	Nichesourcing	Write HS/CS + Paraphrase CS	✓NGO workers, ×non-experts
♠ MULTI-TARGET CONAN (Fanton et al., 2021b)	Hybrid: Nichesourcing and Automated (Human-in-the-loop)	Hybrid: Nichesourcing and Automated (Human-in-the-loop)	Val CS + Edit CS	✓NGO workers, ×academics
♣ DIALOCONAN (Bonaldi et al., 2022)	Hybrid: Nichesourcing and Automated (Human-in-the-loop)	Hybrid: Nichesourcing and Automated (Human-in-the-loop)	Val CS + Edit CS	✓NGO workers
□ MTKGCONAN (Chung et al., 2021c)	Existing dataset (♡)	Automated generation	Ann/Eval CS	✓NGO workers
INTENTCONAN (Gupta et al., 2023)	Existing dataset (♠)	Existing dataset (♠) + Human written	Write CS	×academics
ML-MTCONAN-KN (Bonaldi et al., 2025)	Existing dataset (□)	Human written	Write CS + Edit MT HS/CS	×academics: translators Spanish, Basque, Italian
◊ BENCHMARK (Qian et al., 2019)	Hybrid: Crawling + Crowdsourcing	Crowdsourcing + Automated generation	Val HS + Write CS	×crowdworkers

Table 1: Summary of frequently used existing datasets in counterspeech. The table reports hate speech (HS) and counterspeech (CS) data sources, the type of human input involved in any research stages ('Val': Validating HS/CS instances, 'Ann/Eval': Annotate/Evaluate), and the extent of stakeholder involvement. We list datasets that are used more than twice for both HS and CS sources across the surveyed resources, but exclude those used more than twice for only HS. Note, the 'Hybrid' label is only used when different methods are used within one HS or CS instance; for instance using automated methods to generate CS and then nichesourcing to correct the same CS. The last column gives details about the human involvement with the symbol (✓/×), e.g. row 1 shows that NGO workers are stakeholders given the symbol: ✓.

Include	Exclude
Resources related to human-written counterspeech for dataset creation.	Resources that contain the keyword '*counter-terrorism*' in isolation with none of our other keywords.
Resources related to in-the-wild human-written counterspeech for social media analysis.	Resources with tasks that were irrelevant to the present work, such as *speech-spoofing*.
Resources that do automated counterspeech generation.	Survey resources on counterspeech.

Table 2: Inclusion/exclusion criteria for the review.

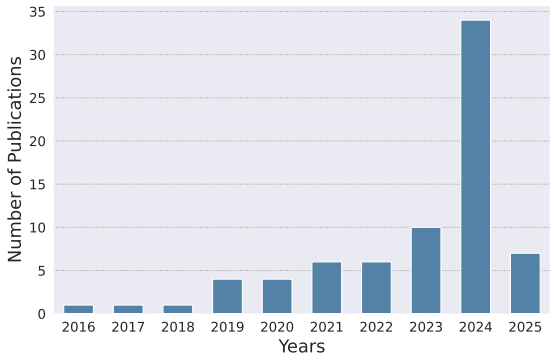

Figure 2: Publications per year up to March 2025.

Eligibility criteria Overall, our goal is to focus on human-written and synthetically generated counterspeech resources in computer science, to answer questions regarding the ways the counterspeech data is sourced, and additionally the level of participatory design involved. Table 2 describes the inclusion and exclusion criteria that were applied. Using these criteria, two of the authors excluded and identified items to review, which were cross-checked by a third author. We then turned our attention to counterspeech resources based on 'in-the-wild' data or performing social media analyses, as these resources may include opinions from experienced users in responding to hate speech online.

Summary of included resources After following the systematic survey process, we were left with 74 items for systematic review that cover wholly or partially automatically generated counterspeech, and the computational analysis of real counterspeech in online settings.

3.1. Results and Discussion

Note: Please see our Github repository for a full record of the surveyed findings, including linguistic metadata about the resources that are not discussed in this work.

Preliminary findings. Figure 2 shows the resources we surveyed by publication year, with a notable recent spike. The results of our survey are given in Table 1, which outlines the most commonly used datasets in counterspeech research and Table 5, which consists of the rest of the surveyed resources. As visually shown in Table 5, close to 50% of the surveyed resources use an existing dataset for sourcing hate speech or counterspeech[5]. Of these resources, as shown in Figure 3 (right), 66% use an iteration of the CONAN (Chung et al., 2019)

[5]Indeed, it was difficult to initially identify whether different resources used the same dataset, given different naming conventions to refer to the same dataset.

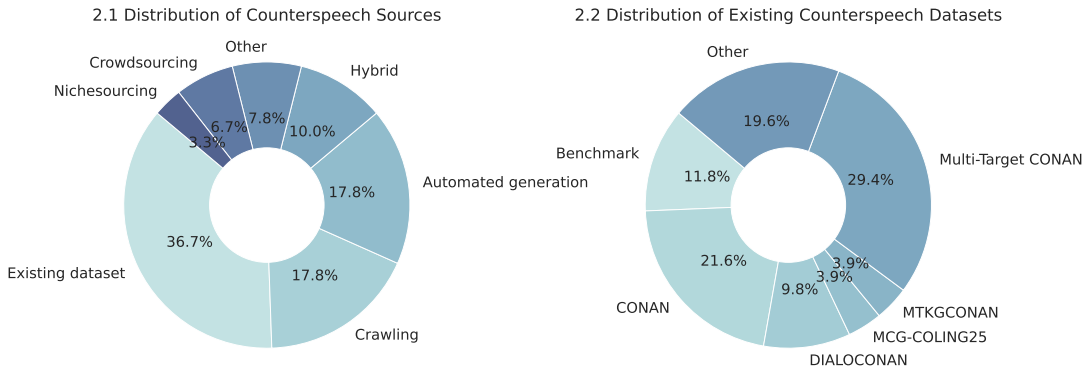

Figure 3: Counterspeech sources and datasets. The percentage reflects the proportion of total sources (N = 88), given that some resources include more than one source.

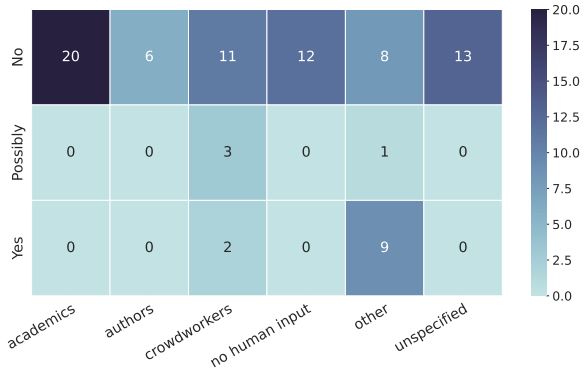

Figure 4: Stakeholder participation. 'Possibly' indicates bystander participation.

datasets, i.e. Multi-Target CONAN (Fanton et al., 2021b), DIALOCONAN (Bonaldi et al., 2022) or MTKGCONAN (Chung et al., 2021c). This is concerning, as constant re-use of these datasets (indeed without benchmarks for comparison and difficulties formulating metrics that capture high-quality counterspeech) can lead to a ceiling effect in terms of performance.

Additionally, the majority of the source datasets were created before LLMs were widely adopted (e.g. CONAN in 2019, Multi-Target CONAN in 2021); these datasets may have been used in the training of proprietary or closed-source models (Balloccu et al., 2024), making it difficult to assess such models fairly for automated counterspeech generation (memorising exact responses to the hate speech, or source datasets containing outdated examples of hate speech)[6]. Figure 3 (left) also shows that 'nichesourcing', or relying on experts to produce responses (Bonaldi et al., 2024a), is the least used method to source counterspeech.

Delgado et al. (2023) give a framework for the extent to which stakeholders can be included in participatory design, from stakeholders *consulting* on participatory design projects as the lowest level of involvement, to then being *included*, *collaborating on*, and finally *ownership* of a project as the highest level of involvement. We analysed the sources for the modes of participatory design according to this framework, to mark six of the resources as *'Consult'*, with an additional 4 as *'Consult/Include(?)'*.

Defining expertise and the value of 'non-expertise' Results show that counterspeech resources use the word *'expert'* in two different ways: for (1) people with relevant specialised experience and knowledge; and (2) NLP academics and students.

Chung et al. (2019); Tekiroğlu et al. (2020); Chung et al. (2021c); Bonaldi et al. (2022); Chung and Bright (2024); Jones et al. (2024) use this term specifically to distinguish NGO workers from non-expert crowdworkers. We also see use of the word *'expert'* when a professional/expert translator is engaged, Chung et al. (2020) for Italian, or Bengoetxea et al. (2024) for Spanish and Basque, and Bonaldi et al. (2025) for all three. However, another group of resources uses this term to indicate domain knowledge in computer science, NLP or linguistics such as in Gupta et al. (2023); Mun et al. (2023); Saha et al. (2024b); Hengle et al. (2024), possibly to distinguish this from data collected from crowdworkers. In Table 5, the latter group can be seen in the column *'Stakeholder involvement'* where we have distinguished between whether the *'experts'* are the authors themselves or other academics with pan-NLP domain expertise. We also use the *'academic'* label when resources don't necessarily claim expert involvement, but do specify academic qualifications as the criteria for annotator recruitment ('3 grad students'). As Figure 4 shows, 26 of the resources we surveyed use either the authors themselves or other academics to annotate or evaluate counterspeech.

Regarding non-experts, some resources may de-

[6]However, this is currently speculative and warrants further research.

liberately use crowdworkers to annotate/evaluate counterspeech, such as in Jones et al. (2024), to get opinions on how difficult it is for an everyday social media user to write counterspeech based on *expert-written NGO guidelines*, and what the barriers are that prevent them from doing so.

Stakeholder and bystander participation It is important to define the terms 'stakeholder' and 'bystander' in order to explain our labelling process in the *'Stakeholder involvement'* column in Table 5. *Stakeholders* refer to agents who practice a niche *'stake'* in interests and processes, such as civil or campaigning gains *[...] "individuals, groups or organisations that share common interests and hold interest in the outcomes of certain decisions or objectives [...]"* (Chidwick et al., 2024). Whilst traditionally referring to business, and often a contested term in feminist research (Wicks et al., 1994), the label is now understood to apply to a range of organisations (Miles, 2017), from policymaking to the third sector. *Bystander* refers to a member of the public and/or community member (who is also a user if referring to internet spaces) who is a first-hand witness to hate speech and holds decision-making power around active and inactive responses, and is a secondary party involved in vicarious trauma.

In our survey, we expand on stakeholder participation to include bystander participation (as shown with the label *'Possibly'* in Table 5). e.g. Lee et al. (2023) recruited annotators with the *explicit requirement* that the annotators have spent time online and encountered hate speech. Ping et al. (2024b); Ding et al. (2024) recruit participants across the US to research (a) why participants may be inclined/disinclined to participate in counterspeech writing online, (b) the frequency with which participants write counterspeech, and (c) participants' opinions on using AI tools to aid in counterspeech writing. While Mun et al. (2024a) utilise both NGO workers and Amazon Mechanical Turk (AMT) workers, there is possible stakeholder participation from (only) the latter, as 94% of the workers reported to have encountered hate speech online and 70% had experience responding to the hate speech. These resources aim for more generalised opinions of bystanders on what are the barriers preventing people from engaging in counterspeech online .

Barriers to participatory design (A lack of) funding and network can create huge barriers to participatory design. While this work focuses on the level of stakeholder involvement in counterspeech resources, we acknowledge these factors as challenging in having such involvement. In one of the surveyed papers, Jones et al. (2024) explain their use of crowdworkers over NGO experts due to *"[...] lack of direct access to expert NGO operators [...]"*. As outlined in Caselli et al. (2021), obtaining funding offers an additional barrier to participatory design research. However it is not simply the funding, but the administrative issues and organisation involved that creates substantial work on those involved in such a project. For example, the data in section 4 took nearly two years to collect after stringent ethical approval from an Institutional Review Board. The process involved extensive research to create a list of organisations that would possibly be of interest and obtaining contacts, securing funding, and organising and facilitating focus groups for those that did respond to us. This is not to mention that dealing with subjects where harm is *"sought"* as a feature of research (e.g. hate speech annotation, oGBV ...) requires carefully planned safeguarding strategies for everyone involved (Kirk et al., 2022).

4. Case study: Addressing Online Gender-Based Violence

While the practices followed by the CONAN datasets centred stakeholder participation, the results of our systematic survey show that this initial goal has been somewhat lost in the resources that followed. An increasing number of datasets reuse the same data with newer algorithmic methods. To understand whether there exist practices used in real world counterspeech that the NLP community is yet to adopt, we conducted a series of structured interactive focus groups (Morgan, 1996) to get stakeholder input on countering hate online, using feminist co-creation and participatory action design practices (Askins, 2018). Our goal is to compile high-level feedback from stakeholders on countering hate online relevant to the community.

We invited oGBV organisations[7] on a country-wide basis. In each focus group, we asked for stakeholder input by deploying open-ended unstructured questions about oGBV into collaborative practical activities (Goessling, 2025). This activity consisted of working with the stakeholders to identify real-world hate-speech samples we collected[8] and get their feedback on the best ways to respond. In the focus groups the authors adopted an observatory and note-taking role, while the stakeholders discussed their insights. At the start of the focus group, we included a high-level explanation of *'AI'*-generated counterspeech, for stakeholders to understand the scope of our project from a computer science perspective. Table 3 gives a brief description of these organisations. Each charity has different specialist focuses, leading to diverse perspectives on counterspeech approaches to oGBV.

[7] Given our specific network of contacts, we decided to focus on the topic of oGBV.

[8] These samples were manually collected mainly from

NGO	Areas of work and expertise
A. EVAW https://www.endviolenceagainst-women.org.uk	A representative collective of violence against women organisations lobbying government for feminist policy on GBV.
B. GLITCH https://glitchcharity.co.uk/	A national charity focused on oGBV especially towards Black women, producing best practice guidance and recommendations for tech companies and government.
C. AMINA https://mwrc.org.uk	A local charity focusing on empowering Muslim and Black & Minority Ethnic (BME) women. Work includes running a helpline to support victims/survivors, providing legal advice regarding immigration concerns and campaigning.
D. SCOTTISH WOMEN'S AID https:// womensaid.scot/	A government-funded charity running advice services for domestic abuse victims. *Note: The NGO worker who participated in this focus group was an expert in financial and online abuse.*
E. COMPASS CENTRE https://www.compasscentre. org/	A small rural GBV charity providing support and advocacy for rape and sexual violence victims/survivors, including a youth group and phone service. *Note: Our focus group specifically engaged with people from the young persons' activist group within this NGO who were survivors of GBV*.

Table 3: NGOs that participated in focus groups to obtain expert insights on countering oGBV.

4.1. Results and Discussion

In section 3, we focused on results from our survey related to participatory design in existing counterspeech research; i.e., which datasets are used, the level and stage of human involvement, terminological discussions around the use of the word '*expert*' to describe annotators, and stakeholder and bystander participation. In this section, we draw on our focus groups with NGOs to interpret and expand on additional survey findings. In particular, we focus on results from our survey that highlight missing elements in current research which would better align with stakeholder-informed feedback. Specifically, aspects of hate speech used to condition counterspeech (a prominent concern among the experts in our focus groups); i.e. missing metadata on the type of hate speech and its targets, lack of sub-categorisation of hate speech, and strategy use in NLP counterspeech. While these results are not discussed in section 3, we elaborate on them here to translate stakeholder feedback into concrete gaps we've identified through our survey. A summary of the feedback from the focus groups can be found in Table 4.

Note: Early on, participants from A used the terms *perpetrator*, *target* and *bystander* to differentiate the roles involved in oGBV, which we adopt.

Focus Issue	Reasoning
Date of HS creation	Interventions are time sensitive, replying to older content can bring further attention towards the HS.
Views and shares of HS	Using these cues to determine if the HS warrants a reply (e.g. weighing benefits between intervening versus prioritising one's own safety).
Reach of the perpetrator	Strategies to adopt differ depending on perpetrator reach.
Use of multiple strategies within the same counterspeech	To answer to different parties involved, i.e. shutting down the perpetrator, providing resources for the target and educating bystanders. Note some of the NGOs had strict policies against engaging the perpetrator.
Sub-category of GBV	Depending on sub-category of GBV (e.g. harassment versus dogpiling), different approaches are adopted.
Anthropomorphism of CS	Weary of bots reinforcing stereotypical 'feminazi' talking points, complications on bots that are explicitly gendered.
Temporality of Language	Perpetrators engage in 'algospeak', finding new ways to escape being flagged by content moderation systems.

Table 4: Summary of key insights from NGOs.

A need for context. Perhaps the starkest difference between counterspeech-focused NLP and stakeholder input was the level of attention given to meta-data pertinent to the hate speech *before* formulating the most appropriate way to respond. Stakeholders considered when the hate speech was created, how often it had been shared and viewed online, asked how many followers the perpetrator has and whether they have a pattern of behaviour in posting such content, and discussed how well the perpetrator seemed to know the target.

Participants from NGOs A and E pointed out that the same hate speech may be shared by a perpetrator with a huge reach online or by a young person in danger of being (further) radicalised, and the strategies they would adopt in those scenarios differ. They favoured sarcasm/shaming to respond to someone with a large following, but adopting a kinder/empathetic tone that would encourage

X/Twitter and included both text and image examples.

Publication	HS source	CS source	Human input and Task (None = ×)	Stakeholder involvement (✓/×) with Details
Tetzlaff et al. (2017)	N/A	Crawling	Val CS	×unspecified
Zubiaga et al. (2024a)	Existing dataset (♡, ♠)	Existing dataset (♡, ♠) + Automated generation	Ann/Eval CS	×unspecified
Ju et al. (2024)	Existing dataset (♣)	Existing dataset (♣) + Automated generation	×	×no human input
Jones et al. (2024)	Existing dataset (♠)	Existing dataset (♠) + Automated generation	Ann/Eval CS	*Possibly*: crowdworkers
Borrelli et al. (2022)	Crawling	Crawling	×	×no human input
Lee et al. (2023)	Existing dataset (♠)	Human annotation	Val CS + Ann/Eval CS	*Possibly*: online 6+ hrs/day
Mathew et al. (2018a)	Crawling	Crawling	Ann/Eval CS	×unspecified
Song et al. (2024)	Crawling	Existing dataset (♡, ♠, ♢) + Crawling	Ann/Eval CS	×academics
Rodrguez et al. (2023)	Existing dataset (□)	Existing dataset (□)	Edit MT HS/CS	×academics
Bengoetxea et al. (2024)	Existing dataset (♡)	Existing dataset (♡)	Edit MT HS/CS + Ann/Eval CS	×professional and native Spanish+Basque
Ping et al. (2024b)	Existing dataset (♠, other)	Crowdsourcing	Write CS + Ann/Eval CS	*Possibly*: crowdworkers
Mun et al. (2023)	Existing dataset (♠, other) + Crawling	Author written + Automated generation	Write CS + Ann/Eval CS	×authors, crowdworkers
Bennie et al. (2025a)	Existing dataset (♠)	Automated generation	×	×no human input
Saha and Srihari (2024a)	Existing dataset (♡, ♣)	Existing dataset (♡, ♣) + Automated generation	Ann/Eval CS	×crowdworkers
Cima et al. (2024)	Crawling	Existing dataset (♠, ♢) + Crawling + Automated generation	Ann/Eval CS	×crowdworkers
Santamaria et al. (2024)	Existing dataset (♣)	Existing dataset (♣) + Automated generation	Ann/Eval CS	×crowdworkers
Garland et al. (2023)	Crawling	Crawling	Val HS/CS	×authors, crowdworkers
Zhang et al. (2024)	Existing dataset (♠, ♢)	Existing dataset (♠, ♢) + Automated generation	Ann/Eval CS	×unspecified
Langer et al. (2019)	Crawling	Crawling	Qualitative analysis CS	×authors
Saha et al. (2022)	Existing dataset (♡, ♢)	Existing dataset (♡, ♢) + Automated generation	Ann/Eval CS	×academics
Garland et al. (2020)	Crawling	Crawling	Ann/Eval CS	×native German crowdworkers
Ding et al. (2024)	Existing dataset (♠, other)	Hybrid	Write CS	*Possibly*: crowdworkers
Mun et al. (2024b)	-	-	Opinons on CS	✓NGO workers, crowdworkers
Saha et al. (2024b)	Existing dataset (other)	Crowdsourcing	Write CS +Ann/Eval CS	×crowdworkers, academics
Hengle et al. (2025)	Existing dataset (other)	Nichesourcing	Ann/Eval CS	× academics
Hassan and Alikhani (2023)	Hybrid	Hybrid + Automated generation	Val HS/CS + Ann/Eval HS/CS + Edit CS	×academics
Song et al. (2025)	Crawling	Crawling	Val CS	×authors
Chung et al. (2021b)	Crawling	Hybrid	Edit CS + Ann/Eval CS	✓NGO workers
Wang et al. (2024a)	Existing dataset (♡, ♠, and □)	Automated generation	×	×no human input
Zhu and Bhat (2021)	Existing dataset (♡, ♢)	Automated generation	Ann/Eval CS	×native English
Tekiroğlu et al. (2020)	Existing dataset (♡, ♢, other)	Hybrid	Val CS + Edit CS	✓NGO workers
Bär et al. (2024)	Crawling	Crawling	×	×no human input
Yu (2022)	Crawling	Crawling	Ann/Eval HS/CS	×crowdworkers
Alyahya and Aldayel (2024)	Existing dataset (♣, other)	Existing dataset (♣, other)	Ann/Eval CS	×crowdworkers
Furman et al. (2023)	Existing dataset (other)	Existing dataset (other)	Ann/Eval CS	×authors, academics
Hickey et al. (2024)	Crawling	Crawling	Ann/Eval CS	×authors, academics
Tonini et al. (2024)	Crawling	Crawling	Val HS/CS + Ann/Eval CS	✓NGO workers, ×academics
Saha and Srihari (2024b)	Existing dataset (♡, ♠, ♣, other)	Existing dataset (♡, ♠, ♣, other)	Ann/Eval CS	×crowdworkers, academics
Wang et al. (2024b)	Existing dataset (♡, ♠)	Existing dataset (♡, ♠)	Ann/Eval CS	×unspecified
Hengle et al. (2024)	Existing dataset (other)	Existing dataset (other)	Ann/Eval CS	×academics
Mathew et al. (2020)	Crawling	Crawling	Ann/Eval HS/CS	×academics
Bonaldi et al. (2024b).	Existing dataset (other)	Automated generation	Ann/Eval CS	×academics
Chung et al. (2020)	Existing dataset (♡)	Existing dataset (♡)	Ann/Eval CS	×native Italian
Zubiaga et al. (2024b)	Existing dataset (other)	Existing dataset (other)	Ann/Eval CS	×unspecified
Lee et al. (2024)	Existing dataset (other)	Existing dataset (other)	×	×no human input
Das et al. (2024)	Crawling	Crowdsourcing	Val HS/CS + Write CS	×academics
Chung et al. (2021a)	Existing dataset (♡)	Existing dataset (♡)	×	×no human input
Gligoric et al. (2024)	Existing dataset (♠, □, other)	Existing dataset (♠, □, other)	Ann/Eval HS/CS	×unspecified
Wadhwa et al. (2024)	Existing dataset (♠)	Existing dataset (other)	×	×no human input
Chung and Bright (2024)	Existing dataset (other) + Hybrid	Hybrid + Automated generation	Write CS + Ann/Eval CS	✓NGO workers, ×crowdworkers
Saha et al. (2024a)	Existing dataset (♡, ♠, ♢)	Existing dataset (♡, ♠, ♢)	×	×no human input
Hong et al. (2024)	Existing dataset (♢, other)	Existing dataset (♢) + Automated generation	Ann/Eval CS	×academics
Rodríguez et al. (2024)	Existing dataset (♠) + Crawling?	Existing dataset (♠) + Nichesourcing?	Edit MT HS/CS + Write CS + Ann/Eval CS	×unspecified
Bennie et al. (2025b)	Existing dataset (other) + Automated generation	Hybrid	Ann/Eval CS + Edit CS	×academics
Furman et al. (2022)	Existing dataset (other)	Crowdsourcing?	Ann/Eval HS + Write CS	×unspecified
Ping et al. (2024a)	Existing dataset (♠, other)	Crowdsourcing	Val HS + Write CS + Ann/Eval CS	✓crowdsourcing + authors
Ziems et al. (2020)	Hybrid + Automated detection	Hybrid + Automated detection	Ann/Eval HS/CS	×academics
Peng and Grimmelmann (2024)	Existing dataset (other)	Existing dataset (other)	Ann/Eval CS	×unspecified
Jiang et al. (2023)	Existing dataset (♠)	Existing dataset (♠)	Ann/Eval CS	×crowdworkers
Saha (2023)	Existing dataset (Unspecified)	Existing dataset (Unspecified)	×	×no human input
Arpinar et al. (2016)			×	×no human input
Alsagheer et al. (2023)	N/A	Crawling	×	×no human input
Mathew et al. (2018b)	Crawling	Crawling	Ann/Eval CS	×academics
Tekiroğlu et al. (2022)	Existing dataset (♠)	Existing dataset (♠)	Ann/Eval CS	×unspecified
Leekha et al. (2024)	Hybrid	Automated generation	Ann/Eval CS	×unspecified
Bonaldi et al. (2023)	Existing dataset (♠)	Existing dataset (♠)	Ann/Eval CS	×unspecified
Halim et al. (2023)	Hybrid: (uses ♡)	Existing dataset (♡)	Ann/Eval CS	×academics

Table 5: Summary of included resources for counterspeech with the same dataset labels and column description from Table 1 (Key: ♡ CONAN, ♠ Multi-target CONAN, ♣ DIALOCONAN, □ MTKGCONAN and ♢ Benchmark)

someone without such a following to reflect on their behaviour, e.g. responding with *'What if this was your your sister?'* NGO B additionally stressed the importance of educational responses to counter oGBV in such cases, pointing out the lack of educational content that addresses young men who feel alienated. NGO A suggested having different strategies

even *within the response* conditioned on different roles, i.e. shutting down/not engaging the perpetrator.[9], providing support or resources for the target and education for the bystanders.

NGOs C and D discussed trends of oGBV in smaller communities and ethnic minority groups; often the perpetrator knows the target personally and will try to socially isolate them from their community by spreading lies or private information (e.g. images) about them. Thus how well the perpetrator knows the target matters; countering targeted harassment will not be the same as countering online bullying or dogpiling. In the community, it is somewhat of a norm to prioritise the metadata of the annotator; providing demographic information such as age, educational background and gender[10].

In contrast, the results of our survey show that NLP counterspeech research does not focus attention on metadata related to the hate speech itself, i.e. it is not present in existing counterspeech datasets and in turn affects research that uses existing datasets (nearly 50% of the resources we surveyed). We also find that $\approx 43\%$ of the resources do not even mention the target group of the hate speech, in particular for those resources using existing datasets. Among the resources that do mention the target, most of them do not consider the information in their design, analysis or evaluation.

While some efforts exist to further sub-categorise GBV in hate speech detection (for instance, *benevolent* vs. *hostile* sexism – see Abercrombie et al. (2023b) for an overview), none of the counterspeech resources including the source datasets in Table 1 have such fine-grained categorisation (e.g. harassment vs. dogpiling) – i.e. it would not be possible to condition counterspeech responses specific to the sub-category as discussed by NGOs C and D. While a recent trend in automated counterspeech generation is to utilise strategies originally proposed by Benesch et al. (2016), these methods are limited by the available linguistic cues present in the hate speech, so strategy generation is not holistic, e.g. considering the audience reach of the perpetrator. Furthermore, to the best of our knowledge, no information on *who* the counterspeech addresses i.e. perpetrator, bystander or target is present in existing resources. Thus NLP counterspeech resources focus on *what* was said in the hate speech given the lack of other metadata available, whereas stakeholders additionally give importance to the surrounding context.

Anthropomorphism. Some issues were raised around the perceived origins of AI-generated counterspeech. Stakeholders from NGO E unanimously agreed that it should be made clear that any counterspeech is artificially generated and not produced by a human. This raises questions of how much store people will put into the responses if they know it is generated by a 'bot'. NGO A discussed being wary of bots reinforcing what are stereotypically considered *'feminazi'* talking points, and that having an anthropomorphically humorous bot is preferable. In the focus group with NGO E, opinion was divided on whether the 'bot' delivering the counterspeech should be explicitly gendered, and if so, how this might impact the effectiveness of its message. There was a consensus that a female persona should not be employed, due to the risk of the message being ignored or diminished as a result. Following this logic, some felt that a male persona would have greater credibility with perpetrators, making them more receptive to the counterspeech message. However, this was objected to by others who felt the bot should strive to be gender neutral – although we note this is difficult to achieve, as people still attribute binary gender to systems despite having minimal gender markers (Aylett et al., 2019; Abercrombie et al., 2023a).

The temporality of language and 'algo-speak'. Resources like datasets encode the context of the period in which the data has been collected and annotated. NGOs A and C brought up that perpetrators often engage in *'algospeak'*, i.e. finding ways to escape being flagged by content moderation algorithms. However, NGO A also stated that perpetrators on newer social media platforms simply repackage oGBV in newer ways; i.e. the implicit nature remains the same.

5. Recommendations

In this section, we distil the results of the focus groups into a practical set of data features that are desirable to collect, which could potentially bridge the gap between how counterspeech is tackled in the real world by stakeholders versus counterspeech-focused NLP.

(AUTOMATICALLY COLLECTED) **Contextual information**, such as **meta-data** from social media (Pérez et al., 2023) (e.g. the number of followers the perpetrator has, how much the hate speech has been viewed and shared) is needed to determine

[9] and noted that some charities have strict policies against engaging the perpetrator.

[10] Demographics have become the norm to provide with paper submissions to ACL, as shown here.

which strategy to adopt. Further **dialogue context** will allow for annotators to make better informed decisions (Sandri et al., 2023). While difficult to determine, it may also reveal information about the connection the perpetrator has to the target (e.g. repeated hate speech within the same dialogue).

(REQUIRING ANNOTATOR EDUCATION) The **sub-category of hate**, for instance, if the sub-category of oGBV is dogpiling, counterspeech generation at scale may be required by prioritising quantity over quality. The **roles**; i.e. paying attention to who is involved and the impact: targets, perpetrators and bystanders. A consensus is emerging that bystander involvement is the key to change. Bystander intervention (Ward) has skyrocketed as a pivotal concept in contemporary GBV studies, where evidence shows that their behavioural decisions, shaped by many socio-cultural and psychological variables (Mainwaring et al., 2023), are key to GBV outcomes, such as prevention, reporting, and harm-reduction.

(REQUIRING STAKEHOLDER INPUT) **Instances of illegal language**, i.e. whether the hate speech contains illegal language and **resources** that educate the bystander and provide support for the target. These may involve working with stakeholders to compile resources on a local level, or consulting stakeholder written sources for up to date factual and educational responses.

6. Conclusions

We systematically reviewed the current state of counterspeech research in NLP. We found that there has been something of a downturn in the extent to which affected stakeholders are engaged in participatory design for this task, with the field heavily relying on a few key datasets and human input limited to a large extent to computer science researchers. To encourage more participatory approaches to NLP counterspeech research, we make recommendations based on feedback from focus groups engaged in tackling real-world hate speech. Our survey reveals that there are many emergent challenges in this field, for instance differing legal jurisdictions for which global definitions (such as the oGBV framework used by the UN and WHO) will not be sufficient. Given that some participants in the focus groups were hesitant about the use of automated counterspeech itself, future work should include reflexive considerations of how hate speech should be tackled online beyond counterspeech – such as better hate speech detection, creation and promotion of tools that individuals can use, or an uptake in human NGO workers or content moderators on social media platforms.

Limitations

This survey focuses exclusively on peer-reviewed NLP and computational social science publications. It does not experimentally validate the impact of stakeholder-informed methods on counterspeech effectiveness. Future research direction requires assessing how such methods for counterspeech could influence the real-world outcomes. Besides, the participatory case study only collaborates with five NGOs with a specific focus on online Gender-Based Violence, which may not fully capture the perspectives of other affected communities, such as religious, or LGBTQ+ groups, etc.

Ethical Statement

This study was approved by our Institutional Review Board (IRB), of the School of Mathematical and Computer Sciences at Heriot-Watt University which reviewed our methodologies and protocols to ensure compliance with ethical standards. Our participatory case study with NGOs was conducted with informed consent, and all participants were made aware of the goals of the research, how their input would be used, and their right to withdraw at any time. Given the sensitive nature of online Gender-Based Violence, we anonymised all identifying details of participants from NGOs, but will release the organisations' names upon acceptance. Furthermore, we compensated the NGOs fairly for their time spent in our focus groups, discussing within our network what is a standard rate for their expertise.

7. Bibliographical References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, Verena Rieser, and Zeerak Talat. 2023a. Mirages. on anthropomorphism in dialogue systems. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4776–4790, Singapore. Association for Computational Linguistics.

Gavin Abercrombie, Aiqi Jiang, Poppy Gerrard-abbott, Ioannis Konstas, and Verena Rieser. 2023b. Resources for automated identification of online gender-based violence: A systematic review. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 170–186, Toronto, Canada. Association for Computational Linguistics.

Dana Alsagheer, Hadi Mansourifar, and W. Shi. 2023. Statistical analysis of counter-hate speech on voice-based social media. pages 1009–1014.

Dana Alsagheer, Hadi Mansourifar, and Weidong Shi. 2022. Counter hate speech in social media: A survey. *arXiv preprint arXiv:2203.03584*.

Ghadi Alyahya and Abeer Aldayel. 2024. Hatred stems from ignorance! Distillation of the persuasion modes in countering conversational hate speech. *ArXiv preprint*, abs/2403.15449.

I. Arpinar, Ugur Kursuncu, and Dilshod Achilov. 2016. Social media analytics to identify and counter Islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online. *2016 International Conference on Collaboration Technologies and Systems (CTS)*, pages 611–612.

Kye Askins. 2018. Feminist geographies and participatory action research: co-producing narratives with people and place. *Gender, Place & Culture*, 25(9):1277–1294.

Matthew P. Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The right kind of unnatural: designing a robot voice. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, CUI '19, New York, NY, USA. Association for Computing Machinery.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Dominik Bär, Abdurahman Maarouf, and Stefan Feuerriegel. 2024. Generative AI may backfire for counterspeech. *ArXiv preprint*, abs/2411.14986.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on Twitter: A field study. *A report for public safety Canada under the Kanishka project*, pages 1–39.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia. ELRA and ICCL.

Michael Bennie, Bushi Xiao, Chryseis Xinyi Liu, Demi Zhang, Jian Meng, and Alayo Tripp. 2025a. CODEOFCONDUCT at multilingual counterspeech generation: A context-aware model for robust counterspeech generation in low-resource languages. *ArXiv preprint*, abs/2501.00713.

Michael Bennie, Demi Zhang, Bushi Xiao, Jing Cao, Chryseis Xinyi Liu, Jian Meng, and Alayo Tripp. 2025b. PANDA - Paired Anti-hate Narratives Dataset from Asia: Using an LLM-as-a-Judge to create the first Chinese counterspeech dataset. *ArXiv preprint*, abs/2501.00697.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA. Association for Computing Machinery.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *Proceedings of the 1st Workshop on Counter-Speech for Online Abuse (CS4OA)*, pages 13–28, Prague, Czechia. Association for Computational Linguistics.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024a. NLP for counter-speech against hate: A survey and how-to guide 🔧. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3480–3499, Mexico City, Mexico. Association for Computational Linguistics.

Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024b. Is safer better? the impact of guardrails on the argumentative strength of LLMs in hate speech countering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3446–3463, Miami, Florida, USA. Association for Computational Linguistics.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Helena Bonaldi, María Estrella Vallecillo-Rodríguez, Irune Zubiaga, Arturo Montejo-Raez, Aitor Soroa, María-Teresa Martín-Valdivia, Marco Guerini, and Rodrigo Agerri. 2025. The first workshop on multilingual counterspeech generation at COLING 2025: Overview of the shared task. In *Proceedings of the First Workshop on Multilingual Counterspeech Generation*, pages 92–107, Abu Dhabi, UAE. Association for Computational Linguistics.

Dario Borrelli, L. Iandoli, J. Ramírez-Márquez, and Carlo Lipizzi. 2022. A quantitative and content-based approach for evaluating the impact of counter narratives on affective polarization in online discussions. *IEEE Transactions on Computational Social Systems*, 9:914–925.

Catherine Buerger and Lucas Wright. 2019. Counterspeech: A literature review. *Available at SSRN 3829816*.

Tommaso Caselli, Roberto Cibin, Costanza Conforti, Enrique Encinas, and Maurizio Teli. 2021. Guiding principles for participatory design-inspired natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 27–35, Online. Association for Computational Linguistics.

Hanna Chidwick, Germaine Tuyisenge, Deborah D DiLiberto, and Lisa Schwartz. 2024. Contradictions and possibilities for change: Exploring stakeholder perspectives of Canada's feminist International Assistance Policy (fiap) and their connection to a future for global health. *PLOS Global Public Health*, 4(11):e0003877.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2024. Understanding counterspeech for online harm mitigation. *Northern European Journal of Language Technology*, 10:30–49.

Yi-Ling Chung and Jonathan Bright. 2024. On the effectiveness of adversarial robustness for abuse mitigation with counterspeech. pages 6988–7002.

Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021c. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it 2020*.

Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, M. Avvenuti, F. Dell'Orletta, and S. Cresci. 2024. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. *ArXiv preprint*, abs/2412.07338.

Mithun Das, Saurabh Kumar Pandey, Shivansh Sethi, Punyajoy Saha, and Animesh Mukherjee. 2024. Low-resource counterspeech generation for Indic languages: The case of Bengali and Hindi. *ArXiv preprint*, abs/2402.07262.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in AI design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23.

Xiaohan Ding, Kaike Ping, Uma Sushmitha Gunturi, Buse Çarik, Sophia Stil, Lance T. Wilhelm, T. Daryanto, James Hawdon, Sang Won Lee, and Eugenia H. Rho. 2024. CounterQuill: Investigating the potential of human-AI collaboration in online counterspeech writing. *ArXiv preprint*, abs/2410.03032.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3226–3240, Online. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

D. Furman, Pablo Torres, José A. Rodríguez, Diego Letzen, María Vanina Martínez, and Laura Alonso Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. pages 2942–2956.

D. Furman, Pablo E. Torres, José Raúl Rodríguez Rodríguez, Lautaro Martínez, L. A. Alemany, Diego Letzen, and María Vanina Martínez. 2022. Parsimonious argument annotations for hate speech counter-narratives. *ArXiv preprint*, abs/2208.01099.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and M. Galesic. 2023. Correction: Impact and dynamics of hate and counter speech online. *EPJ Data Science*, 12:1.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. NLP systems that can't tell use from mention censor counterspeech, but teaching the distinction helps. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.

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

Kristen P. Goessling. 2025. Learning from feminist participatory action research: A framework for responsive and generative research practices with young people. *Action Research*, 23(1):48–70.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada. Association for Computational Linguistics.

Sadaf Md. Halim, Saquib Irtiza, Yibo Hu, L. Khan, and B. Thuraisingham. 2023. WokeGPT: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10.

Sabit Hassan and Malihe Alikhani. 2023. Discgen: A framework for discourse-informed counterspeech generation. pages 420–429.

Amey Hengle, Aswini Kumar, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. CSEval: Towards automated, multi-dimensional, and reference-free counterspeech evaluation using auto-calibrated LLMs. *ArXiv preprint*, abs/2501.17581.

Amey Hengle, Aswini Kumar, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakroborty. 2024. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with RLAIF. *ArXiv preprint*, abs/2403.10088.

Daniel Hickey, Matheus Schmitz, D. Fessler, P. Smaldino, Kristina Lerman, Goran Muri'c, and Keith Burghardt. 2024. Hostile counterspeech drives users from hate subreddits. *ArXiv preprint*, abs/2405.18374.

Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-constrained large language models for countering hate speech. pages 4523–4536.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, and Wenxian Wang. 2023. ReZG: Retrieval-augmented zero-shot counter narrative generation for hate speech.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Zhuoya Ju, Haiyang Wang, Qiang Li, Wenhua Liu, Jiaqi Han, and Ping Li. 2024. A multi-agent parallel management method for decision-making in countering hate speech. *2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 604–608.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Amanda Langer, M. Kaufhold, Elena Runft, Christian Reuter, Margarita Grinko, and V. Pipek. 2019. Counter narratives in social media: An empirical study on combat and prevention of terrorism.

Seungyoon Lee, Dahyun Jung, Chanjun Park, Seolhwa Lee, and Heu-Jeoung Lim. 2023. Alternative speech: Complementary method to counter-narrative for better discourse. *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1438–1442.

Seungyoon Lee, Chanjun Park, Dahyun Jung, Hyeonseok Moon, Jaehyung Seo, Sugyeong Eo, and Heu-Jeoung Lim. 2024. Leveraging pre-existing resources for data-efficient counter-narrative generation in Korean. pages 10380–10392.

R. Leekha, Olga Simek, and Charlie Dagli. 2024. War of words: Harnessing the potential of large language models and retrieval augmented generation to classify, counter and diffuse hate speech. *The International FLAIRS Conference Proceedings*.

Chelsea Mainwaring, Fiona Gabbert, and Adrian J Scott. 2023. A systematic review exploring variables related to bystander intervention in sexual violence contexts. *Trauma, Violence, & Abuse*, 24(3):1727–1742.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on Twitter. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018a. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, S. Maity, Pawan Goyal, and Animesh Mukherjee. 2018b. Thou shalt not hate: Countering online hate speech. pages 369–380.

S Miles. 2017. Stakeholder theory classification, definitions and essential contestability. 21–47.

David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G. Altman. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4):264–269. PMID: 19622511.

Nyalleng Moorosi, Raesetje Sefala, and Sasha Luccioni. 2023. AI for whom? Shedding critical light on AI for social good. In *NeurIPS 2023 Computational Sustainability: Promises and Pitfalls from Theory to Deployment*.

David L. Morgan. 1996. Focus groups. *Annual Review of Sociology*, 22(Volume 22, 1996):129–152.

Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Communications of the ACM*, 36(6):24–28.

Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. 2023. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. pages 9759–9777.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024a. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Jimin Mun, Cathy Buerger, Jenny T. Liang, Joshua Garland, and Maarten Sap. 2024b. *Counterspeakers' Perspectives: Unveiling Barriers and AI Needs in the Fight against Online Hate*.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120.

Kenny Peng and James Grimmelmann. 2024. Rescuing counterspeech: A bridging-based approach to combating misinformation. *ArXiv preprint*, abs/2410.12699.

Kaike Ping, James Hawdon, and Eugenia H. Rho. 2024a. Perceiving and countering hate: The role of identity in online responses. *ArXiv preprint*, abs/2411.01675.

Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia H. Rho. 2024b. Behind the counter: Exploring the motivations and barriers of online counterspeech writing. *ArXiv preprint*, abs/2403.17116.

Juan Manuel Pérez, Franco M. Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, and Viviana Cotik. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

María Estrella Vallecillo Rodrguez, Arturo Montejo-Ráez, and M. T. M. Valdivia. 2023. Automatic counter-narrative generation for hate speech in Spanish. *Proces. del Leng. Natural*, 71:227–245.

María Estrella Vallecillo Rodríguez, Maria Victoria Cantero Romero, Isabel Cabrera De Castro, L. A. U. López, Arturo Montejo Ráez, and M. Martín-Valdivia. 2024. Overview of RefutES at IberLEF 2024: Automatic generation of counter speech in Spanish. *Proces. del Leng. Natural*, 73:449–459.

Punyajoy Saha. 2023. Self-supervision and controlling techniques to improve counter speech generation. *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024a. On zero-shot counterspeech generation by LLMs. pages 12443–12454.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024b. CrowdCounter: A benchmark type-specific multi-target counterspeech dataset. *ArXiv preprint*, abs/2410.01400.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech. *ArXiv preprint*, abs/2205.04304.

Sougata Saha and R. Srihari. 2024a. Consolidating strategies for countering hate speech using persuasive dialogues. *ArXiv preprint*, abs/2401.07810.

Sougata Saha and R. Srihari. 2024b. Integrating argumentation and hate-speech-based techniques for countering misinformation. pages 11109–11124.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.

Selene Baez Santamaria, Helena Gómez-Adorno, and Ilia Markov. 2024. Contextualized graph representations for generating counter-narratives against hate speech. pages 7664–7674.

Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2024. Assessing the human likeness of AI-generated counterspeech. pages 3547–3559.

Xiaoying Song, Sharon Lisseth Perez, Xinchen Yu, Eduardo Blanco, and Lingzi Hong. 2025. Echoes of discord: Forecasting hater reactions to counterspeech. *ArXiv preprint*, abs/2501.16235.

Francesca Stevens, Florence E. Enock, Tvesha Sippy, Jonathan Bright, Miranda Cross, Pica Johansson, Judy Wajcman, and Helen Z. Margetts. 2024. Women are less comfortable expressing opinions online than men and report heightened fears for safety: Surveying gender differences in experiences of online harms.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pretrained language models for producing counter narratives against hate speech: a comparative study. pages 3099–3114.

Emily J. Tetzlaff, E. Jago, Ann Pegoraro, and T. Eger. 2017. *#DistractinglySexy: How Social Media was used as a Counter Narrative on Gender in STEM*.

Marcus Tomalin and Stefanie Ullmann, editors. 2023. *Counterspeech. Multidisciplinary Perspectives on Countering Dangerous Speech*. Taylor & Francis.

Vittoria Tonini, Simona Frenda, M. Stranisci, and Viviana Patti. 2024. How do we counter hate speech in Italy?

Sahil Wadhwa, Chengtian Xu, Haoming Chen, Aakash Mahalingam, Akankshya Kar, and Divya Chaudhary. 2024. Northeastern Uni at multilingual counterspeech generation: Enhancing counter speech generation with LLM alignment through direct preference optimization. *ArXiv preprint*, abs/2412.15453.

Haiyang Wang, Yuchen Pan, Xin Song, Xuechen Zhao, Minghao Hu, and Bin Zhou. 2024a. F^2RL: Factuality and faithfulness reinforcement learning framework for claim-guided evidence-supported counterspeech generation. pages 4457–4470.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024b. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided LLMs. pages 9131–9142.

Jeanne Ward. Risks and opportunities for adopting 'bystander intervention approaches' to discourage, prevent or interrupt gender-based violence in humanitarian settings. *Gender-Based Violence AoR*.

Andrew C. Wicks, Daniel R. Gilbert, and R. Edward Freeman. 1994. A feminist reinterpretation of the stakeholder concept. *Business Ethics Quarterly*, 4(4):475–497.

Xinchen Yu. 2022. Hate speech and counter speech detection: Conversational context does matter. pages 5918–5930.

Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, and Xian Sun. 2024. COT: A generative approach for hate speech counter-narratives via contrastive optimal transport. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 9:740–756.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2020. Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

I. Zubiaga, A. Soroa, and R. Agerri. 2024a. A LLM-based ranking method for the evaluation of automatic counter-narrative generation. pages 9572–9585.

I. Zubiaga, A. Soroa, and Rodrigo Agerri. 2024b. Ixa at refutES 2024: Leveraging language models for counter narrative generation.