Counterbalancing Hate with Positivity: A Survey of Counterspeech

Anonymous ACL submission

Abstract

While the internet and social web platforms provide an effective stage to share ideas, exchange opinions, and debate issues at an unprecedented scale, ensuring civil discourse remains a major concern. Counterspeech is a broad umbrella term used for content that constitutes a response that takes issues with hateful, harmful, or extremist content. Counterspeech presents an appealing path to counter hateful or extreme content without requiring the removal of such content thus better serving free speech. This survey presents a broad take on counterspeech, outlining operationalization subtleties, pointing to existing resources to train supervised solutions, listing cutting-edge generative AI efforts, and finally, discussing extant gaps and challenges informing future research.

1 Introduction

If there be time to expose through discussion the falsehood and fallacies, to avert the evil by the processes of education, the remedy to be applied is more speech, not enforced silence.
– Louis Brandeis (Whitney v California, 1927, U.S. Supreme Court, p. 377).

Ensuring civil discourse in social media platforms remains a critical challenge (Saha et al., 2023). Beyond creating an unsafe web space for a wide range of minority groups, hate speech may have far-reaching "off-line" consequences such as negatively impacting the mental and physical well-being of targeted entities and triggering physical violence (Müller and Schwarz, 2021).

Major online platforms have adopted a myriad of strategies to combat online toxicity that include detection and subsequent deletion of objectionable content, de-platforming (Agarwal et al., 2022), and shadowbanning users (Zeng and Kaye, 2022). However, such mitigation strategies may not only suppress illicit speech, but also muffle valuable speech (Bamman et al., 2012; Duffy and Meisner, 2023) and can be seen as directly conflicting with free speech, a philosophy deeply cherished by several democracies. Also, such strategies primarily disperse hate speech rather than reduce it (Chandrasekharan et al., 2017; Horta Ribeiro et al., 2023). In contrast, *counterspeech* (Benesch et al., 2016) presents a viable alternative that explores adding (or highlighting) more speech as a remedial measure (Strossen, 2018). This survey presents a detailed exposition of the counterspeech literature that has considerably grown over the last decade.

Motivation: According to Benesch et al., 2016, counterspeech is defined as "*a response that takes issue with hateful, harmful, or extremist content*". An illustrative example is presented in Table 1. The advantages of counterspeech over traditional hate speech moderation strategies are the following. First, traditional hate speech moderation requires authority in some form or shape (e.g., government, platform owners, moderators of a forum). In contrast, civic participation in counterspeech has little or no barrier – it can be practiced by anyone. Second, the removal of content or blocking of users typically requires a high bar to meet, platforms and governments have varying standards. While overt hate speech may qualify for moderation through removal, covert hate speech or microaggressions seldom get removed (Baider, 2023). Finally, as championed by Supreme Court Justice Louis Brandeis in his landmark judgment on Whitney v California, using *more speech* to counterspeech aligns better with the philosophy of upholding free speech.

How the landscape has evolved: Over the last few years, growing political polarization (Demszky et al., 2019; KhudaBukhsh et al., 2021), xenophobia (Kende and Krekó, 2020), and conflicts caused by international disputes (Palakodety et al., 2020a; Pronoza et al., 2021; Tyagi et al., 2020) have spilled their negativity into a volatile online world. Civic participation in ensuring civil discourse has never been so essential. With generative

AI in the mix, where bad actors have access to ready resources that can create unbridled toxicity at scale, counterspeech (both human-generated and machine-assisted) can be a vital tool to keep the information ecosystem safe (Hartvigsen et al., 2022; Dutta et al., 2024).

The present survey: While this survey covers the rich literature that directly uses the technical term *counterspeech* or *counter narrative* in their contributions, a broad range of studies capture the spirit of counterspeech without explicitly mentioning the term or often coining new terms (e.g., *hope speech* (Palakodety et al., 2020a) and *help speech* (Palakodety et al., 2020b)). The goal of our survey is to present a comprehensive picture of this broad notion of counterspeech with a detailed outline of extant gaps and challenges.

Hate speech (HS)	*The world would be a better place without Muslims. They are only killing and raping our children.*
Counterspeech ($\mathcal{CS}$)	On the contrary, most children abuse is operated by people they know: a relative, family friends, sports coach, someone in a position of trust and authority. Besides, Muslims help people - A Muslim woman rushed to help the victims of a triple stabbing in Manchester on New Year's Eve.

Table 1: An example of counterspeech in response to hate speech (taken from Chung et al., 2021b) employing the counterspeech strategy ***present facts*** as outlined in Benesch et al., 2016.

2 Definitions and operationalization

Source	Definition
(Benesch et al., 2016)	[Counterspeech can be defined as] *a response that takes issue with hateful, harmful, or extremist content.*
(Bartlett and Krasodomski-Jones, 2016)	[Counterspeech can be defined as a] *crowd-sourced response to extremism or hateful content*
(Saltman and Russell, 2014)	[Counterspeech can be defined as] *any articles, videos, speeches and other material that seeks to challenge hateful and/or extreme views through positive messaging and narratives*
(Bahador, 2021)	[Counterspeech can be defined as] *communication that directly responds to the creation and dissemination of hate speech with the goal of reducing harmful effects.*

Table 2: Definitions of $\mathcal{CS}$ found in the literature.

Definitions: Multiple definitions of counterspeech ($\mathcal{CS}$[1]) exist in the literature with substantial thematic overlap. Table 2 lists a few illustrative, well-known definitions. This survey treats the term *counterspeech* as a hypernym, broadly encompassing all these definitions as well as considering other speech definitions with a narrower focus but with similar intent. For instance, $\mathcal{CS}$ can also take the form of an alternative narrative where positive stories about social values, tolerance, openness, freedom, and democracy can be used to counter extreme or hateful content. Certain scholarly works focused on specific exogenous shocks share a similar philosophy. For instance, Palakodety et al., 2020a introduce *hope speech* as de-escalating, user-generated, social media content in the context of online discourse amid international conflicts. While *Hope speech* was coined in the context of the 2019 India-Pakistan Pulwama conflict, Chakravarthi et al., 2022 expanded this term to a broader scope of speech championing equality, diversity, and inclusion.

Takeaways

The broad takeaway from this section is $\mathcal{CS}$ is a nuanced concept and there is no one-size-fits-all operationalization of counterspeech. Different researchers approach this challenging task with subtle variations. As we observe work grounded in psychology (Mun et al., 2023), a deeper connection with other social science fields will benefit the operationalization of $\mathcal{CS}$ through adding strategies yet to be included in computational frameworks. For instance, *bending* as proposed by Caponetto and Cepollaro, 2023. Bending is defined as a strategy where the original intent of the speaker is distracted through a clever response that deliberately assumes that the offender's intent is different. A better integration with ethics, and social psychology literature will thus enrich $\mathcal{CS}$ research.

Operationalizing counterspeech: Much of the existing $\mathcal{CS}$ datasets and $\mathcal{CS}$ literature (Tuck and Silverman, 2016; Mathew et al., 2019; Chung et al., 2019) rely on the framework presented by Benesch et al., 2016 that outlines the following eight strategies for $\mathcal{CS}$: (i) ***present facts***: fact-based correction of misstatements or misperceptions, (ii) ***point out hypocrisy***: pointing out hypocrisy or contradic-

[1]We use the shorthand $\mathcal{CS}$ for counterspeech interchangeably.

tions, (iii) ***warn about consequences***: warning of possible offline and online consequences of speech (iv) ***claim affiliation***: identification with the original speaker or target group (v) ***denounce***: denouncing speech as hateful or dangerous (vi) ***use media***: use of media (vii) ***use humor***: use of humor or sarcasm (viii) ***use a particular tone***: use of a particular tone, e.g. an empathetic one. Certain studies including (Mathew et al., 2019; Chung et al., 2019; Baider, 2023; Gupta et al., 2023; Mun et al., 2023) introduce minor differences although the central theme revolves around those stated above.

Detailed operationalizable definitions of *hope speech*, *help speech* and *supportive speech* with examples can be found in (Palakodety et al., 2020a,b; Yoo et al., 2021).

3 Counterspeech datasets

Existing literature adopts diverse strategies and varying standards to identify $\mathcal{CS}$ for dataset curation. A dataset summary follows next.

Monolingual datasets: Wright et al., 2017 describe "golden conversations" as a three-step exchange where hate speech posted in the first step is countered in the second step and the third step indicates a favorable impact on the account that posted the hate speech (e.g., an apology, recanting, or deleting the content, etc.). The third step, as noted in several studies, is elusive (Mathew et al., 2019). Connecting $\mathcal{CS}$ with the hate speech to which it is responding, can also be tricky. For example, YouTube provides a two-stage hierarchy of comments and replies. Even then, users intending to reply to a specific comment often post a standalone comment instead. Mathew et al., 2019 circumvent this challenge by studying responses to hateful YouTube videos. Garland et al., 2020 study the interaction between two opposing groups *Reconquista Germanica* and *Reconquista Internet*. Other common strategies include studying datasets known to trigger hate speech (Baider, 2023), tracking hashtags (Yoo et al., 2021), and regular expressions (Mathew et al., 2019). Most of these datasets vary along data collection method, dissemination protocol, rater diversity, target groups, and language. Citing real-world social web data not meeting the desired $\mathcal{CS}$ standards, Chung et al., 2019 presents a dataset where the counter responses are written by experienced NGO operators. Human-in-the-loop settings, where machine-generated responses are shortlisted and edited by humans, are adopted in (Fanton et al., 2021; Tekiroglu et al., 2022). CrowdCounter (Saha et al., 2024) is a novel dataset of 3,425 hate speech–counterspeech pairs across six counterspeech types (empathy, humor, questioning, warning, shaming, contradiction) where the annotation platform is designed to encourage annotators to generate type-specific, diverse, and high-quality counterspeech responses, ensuring a more effective and well-structured dataset.

Multilingual datasets: While significant progress has been made in developing automated counterspeech models, much of this research has been centered on English. Addressing this gap, recent studies have explored multilingual approaches to counterspeech, particularly for low-resource languages. (Das et al., 2024) proposed a dataset comprising 5,062 abusive speech/counterspeech pairs (2,460 in Bengali and 2,602 in Hindi) where the authors stated that for automated generation of counterspeech monolingual models outperform interlingual transfer setups. Unlike the former dataset where counterspeeches were annotated manually, (Bengoetxea et al., 2024) created a dataset, developed via machine translation and professional post-editing, aligns with the English CONAN dataset. Results indicate that manually revised data significantly improves the quality of counterspeech generation compared to silver-standard machine-translated data. In addition, cross-lingual data augmentation benefits structurally similar languages (e.g., Spanish and English) but proves less effective for language isolates like Basque.

Transience: Another key challenge to curating social web datasets is the transience of the social web. To protect users' right to be forgotten, many research groups disseminate only social web identifiers (e.g., tweet id, or YouTube comment id) that need to be rehydrated later. While this style of dissemination ensures that if a social web post is removed from the Internet, the content is automatically deleted from the dataset as well, it poses reproducibility challenges. Chung et al., 2019 report that more than 60% of the dataset presented by Mathew et al., 2020 became irretrievable and presented a copy-free dataset as a remedial measure.

$\mathcal{CS}$ in real-world conversations is rare. Palakodety et al., 2020a estimate 2.5% of the social media discourse authored in English was *hope speech*. Similar to $\mathcal{CS}$, *hope speech* and *help speech* also have nuanced definitions with sub-categories. Palakodety et al., 2020b present

a semantic sampling approach where for each sub-category seed examples are chosen. These seed examples either come from the dataset or are provided by the annotators. Next, a nearest-neighbor sampling in the document embedding space identifies a smaller subset which is subsequently labeled. Similar semantic sampling approaches with richer embeddings provided by cutting-edge large language models may alleviate the data scarcity problem often encountered in $\mathcal{CS}$ research relying on real-world data.

Takeaways

We identify the following key gaps in $\mathcal{CS}$ datasets: (1) lack of focus on annotation diversity and transparency in annotator demographics reporting; (2) lack of connection with annotation subjectivity literature and participatory AI frameworks; and (3) narrow coverage addressing a small set of target groups and languages. Section 6 presents a detailed exposition on these gaps.

4 $\mathcal{CS}$ detection and generation

Detection: $\mathcal{CS}$ detection is effectively a stance classification task. The literature on $\mathcal{CS}$ detection methods follows the trajectory of machine learning methods' advancements with earlier approaches relying on support vector machines (e.g. Palakodety et al., 2020a,b), boosting methods (e.g. Mathew et al., 2019) embedding-based methods, to the next generation approaches relying on `BERT`-based methods (e.g. Vidgen et al., 2020; Goffredo et al., 2022), and `XLNet` (e.g. (Vidgen et al., 2020)). In what follows, we highlight a few papers with interesting information science insights. Garland et al., 2020 developed a two-stage classification process using paragraph embeddings and classification via regularized logistic regression. This method was augmented by an ensemble learning approach, leveraging a "panel of experts" to enhance classification accuracy. Yu et al. argued for the importance of considering the preceding comment (context) to accurately classify a comment as hate speech, $\mathcal{CS}$, or neutral. Testing both context-unaware and context-aware models, the study demonstrated superior performance for the latter.

Among the prominent research in detecting $\mathcal{CS}$ in non-English based languages, Goffredo *et al.* (Goffredo et al., 2022) focused on detecting $\mathcal{CS}$ in Italian, while Chung et al., 2021a explored $\mathcal{CS}$ detection in a multilingual setting in English, French, and Italian.

Takeaways

We observe two key ways the detection research can be further strengthened: (1) incorporating rigorous in-the-wild evaluations, and (2) testing robust generalizability.

Generation: Automated $\mathcal{CS}$ generation leverages generative AI to counter hate speech (Qian et al., 2019; Padhi et al., 2025). Tekiroglu et al., 2022 suggest that among the proposed methodologies for automatically generating $\mathcal{CS}$, the use of autoregressive models combined with stochastic decoding mechanisms performs the best in producing novel, diverse, and informative $\mathcal{CS}$. On the other hand, Bennie et al., 2025 proposed the CODEOFCONDUCT model, a context-aware model fine-tuned on multilingual datasets which demonstrated state-of-the-art performance in the MCG-COLING-2025 shared task. It ranked first for Basque, second for Italian, and third for English and Spanish. The model's effectiveness in low-resource languages underscores the potential of fine-tuned multilingual architectures for $\mathcal{CS}$ generation. Two central themes emerge from the recent lines of work. $\mathcal{CS}$ generation for – (a) ***catering to different types of hate speech*** and (b) ***catering to different nuances in generation***. The first line of work concentrates on generating $\mathcal{CS}$ based on subtle variations and nuances in offensive content. Ashida and Komachi, 2022 extended counter-narrative generation targets to include microaggressions and demonstrates the effectiveness of few-shot prompting with the use of pre-trained large language models (LLMs) such as `GPT-2`, `GPT-Neo`, and `GPT-3` for generating $\mathcal{CS}$. Lee et al., 2022 introduced the first dataset, dubbed `ELF22`, for countering online trolls, labeled with counter strategies and detailed contextual information, which facilitates the automatic generation of counter-responses. By categorizing trolls into overt and covert types and annotating counter-responses with seven distinct strategies, `ELF22` enables the development of models capable of generating responses based on the context and intended strategy.

$\mathcal{CS}$ generation approaches that consider nuances in generation strategies fall into the second category. Doğanç and Markov, 2023 introduced the concept of incorporating author profiling aspects, such as age and gender, into `GPT-2` and `GPT-3.5` models to enhance $\mathcal{CS}$ personalization.

Gupta et al., 2023 proposed QUARC, a two-stage framework for generating intent-conditioned $\mathcal{CS}$ whereas Saha et al., 2022 presented an ensemble of generative discriminators (GEDI) to guide the generation of a DialoGPT model towards generating $\mathcal{CS}$ that is more polite, detoxified, and emotionally resonant. To enhance diversity and relevance in $\mathcal{CS}$ generation, Zhu and Bhat, 2021 first generated a diverse pool of $\mathcal{CS}$ candidates, then pruned ungrammatical ones using a BERT model, and finally selected the most relevant response through a novel retrieval-based method. Saha et al., 2024 recently introduced the CrowdCounter dataset consisting of hate speech–counterspeech pairs across six $\mathcal{CS}$ types (empathy, humor, questioning, warning, shaming, contradiction) and used this dataset to fine-tune multiple LLMs to generate type-specific $\mathcal{CS}$ that is more effective.

Takeaways

We find multiple gaps in $\mathcal{CS}$ generation schemes. (i) First, most generative models are for English & knowledge transfer mechanisms like context distillation may be used for generation in other languages. (ii) New safety training & alignment methods like preference modelling (Huang et al., 2023), task arithmetic (Ortiz-Jimenez et al., 2023), model arithmetic (Banerjee et al., 2025) etc. which can be used to generate $\mathcal{CS}$ that is more aligned with user preference.

Several lines of recent research delve deeper into the underlying structure of hate speech and $\mathcal{CS}$. For instance, Mun et al., 2023 presented an innovative approach to $\mathcal{CS}$, focusing on dispelling the underlying stereotypes and biases implied in hateful language by identifying and evaluating six psychologically inspired strategies to counteract stereotypical implications, moving beyond simple denouncement of hate speech. Similarly, Saha and Srihari, 2024 control the generation of counterarguments by leveraging features based on argument structure and reasoning (using Walton argument schemes), counter-argument speech acts, and human characteristics (such as Big-5 personality traits and human values). Hassan and Alikhani, 2023 proposed DisCGen, which includes a taxonomy of $\mathcal{CS}$ derived from Segmented Discourse Representation Theory (SDRT) and discourse-informed prompting strategies for generating contextually relevant $\mathcal{CS}$ with LLMs. Recent approaches have tackled three recurrent themes in $\mathcal{CS}$ generation: data sparsity, diversity, and relevance in different ways that include leveraging external knowledge to produce more informative and contextually relevant responses (Chung et al., 2021b); utilization of RAG-based methods (Jiang et al., 2023); and regularization methods (Bonaldi et al., 2023).

Generation cum alignment: Halim et al., 2023 addressed this challenge through retraining language models with their WokeCorpus, augmenting Hatespeech-$\mathcal{CS}$ pair by generating synthetic $\mathcal{CS}$, and subsequently fine-tuning the model to generate $\mathcal{CS}$. To address the challenge of insufficient $\mathcal{CS}$ in multilingual contexts, Chung et al., 2020 pioneer the use of neural machine translation systems to generate "silver data" from various languages. This data is then utilized to refine GePpeTto, a model based on GPT-2 and customized for Italian. A key finding from their research is the effectiveness of integrating translated (silver) data with original (gold) data in $\mathcal{CS}$ generation. CoARL (Hengle et al., 2024) is a novel framework designed to enhance $\mathcal{CS}$ generation by capturing the pragmatic implications of social biases embedded in hateful statements. It follows a three-phase approach: (1) sequential multi-instruction tuning, where the model learns to recognize intents, reactions, and harms associated with offensive statements; (2) task-specific adaptation, where low-rank adapter weights are trained to generate intent-conditioned $\mathcal{CS}$; and (3) reinforcement learning, which fine-tunes the generated responses for effectiveness and non-toxicity. Extensive human evaluations further confirm CoARL's superiority in generating more contextually appropriate and impactful $\mathcal{CS}$ compared to existing models, including leading LLMs like ChatGPT. Similarly, DART (Wang et al., 2024) is a novel LLM-based framework for $\mathcal{CS}$ generation, leveraging a DuAl-discRiminaTor mechanism to guide the decoding process. It incorporates two key discriminators: an intent-aware discriminator to ensure responses align with specific conversational intents and a hate-mitigating discriminator to enhance the effectiveness of $\mathcal{CS}$ in reducing hate. Finally, Hong et al., 2024 generated $\mathcal{CS}$ guided by specific conversational outcomes and evaluated their effectiveness. They integrate two key conversational objectives into the text generation process by LLMs: reducing conversation incivility

and encouraging non-hateful reentry by haters. To achieve this, they used three approaches: instruction prompting, LLM fine-tuning, and reinforcement learning (RL) for LLMs. They demonstrated that these methods successfully steer $\mathcal{CS}$ generation toward the desired outcomes.

5 Field experiments and surveys

Is $\mathcal{CS}$ effective in practice? Early research on $\mathcal{CS}$ was inspired by real-world observations, albeit anecdotal, of $\mathcal{CS}$ influencing online harm mitigation. Wright et al., 2017 recount that the earliest observation of real-world effectiveness of $\mathcal{CS}$ was observed in 2013 as part of a dangerous speech project in a Twitter study following the presidential election in Kenya. Manual inspection in Mathew et al., 2020 revealed a user who later apologized to everyone saying that they were sorry for what they did. However, Mathew et al., 2020 noted that such public apologies were indeed rare.

What are the barriers in $\mathcal{CS}$ writing? Current research (Ping et al., 2024) has primarily focused on the attributes and impact of online $\mathcal{CS}$, leaving a gap in understanding who participates in $\mathcal{CS}$ and what factors influence their engagement or deterrence. To address this, a survey was conducted with 458 English-speaking US participants analyzing and key motivations and barriers to online $\mathcal{CS}$ engagement. The findings indicate that having been a target of online hate serves as a significant driver of frequent $\mathcal{CS}$ engagement. Younger individuals, women, those with higher education levels, and regular witnesses to online hate were found to be more reluctant to engage in $\mathcal{CS}$ due to concerns about public exposure, retaliation, and third-party harassment. In addition, participants' willingness to use AI technologies, such as ChatGPT, for $\mathcal{CS}$ writing was explored. A similar study (Mun et al., 2024) investigated the barriers to using AI in scaling up $\mathcal{CS}$ efforts through in-depth interviews with 10 highly experienced counterspeakers, along with a large-scale public survey involving 342 everyday social media users. From participant responses, four main types of barriers related to resources, training, impact, and personal harms were identified. Further, overarching concerns regarding authenticity, agency, and functionality in the use of AI tools for $\mathcal{CS}$ were also revealed.

How to evaluate the effectiveness of $\mathcal{CS}$? Different $\mathcal{CS}$ strategies may have varying success depending on the target and the nature of hate. In an experimental design, Hangartner et al., 2021 identified 1,350 Twitter accounts that posted racist or xenophobic tweets. The treatment group was countered with *empathy*, *warning of consequences*, and *humor*. The study revealed that empathy had a noticeable, positive impact on the treatment group in increasing retrospective deletion of xenophobic and racist messages while the other two strategies (warning of consequences and humor) did not have any discernible impact. While *warning of consequences* and *humor* had no observable effect on retrospective deletion in this study, mathew2019thou observed that humour was particularly useful in countering hate against the LGBTQ+ community. Beyond the behavioral shift of the offender, effectiveness can also be evaluated through the lens of bystanders' attitudinal shifts. A survey experiment on 1,250 YouTube users in South Korea reveals complex gender dynamics. The study suggests that users are more likely to report sexist hate speech if the $\mathcal{CS}$ is posted by a female than a male user (Kim et al., 2023). However, the study shows a curious pattern that if the $\mathcal{CS}$ posted by a female gets considerable upvotes, males are less likely to report possibly due to the elevated status of the $\mathcal{CS}$. However, Van Houtven et al., 2024 present that in case of racism, $\mathcal{CS}$ initiated by a majority group member had better success in mitigating hate than a minority group member.

Takeaways

While field studies and surveys can provide evidence of efficacy, robust assessment of the viability of $\mathcal{CS}$ as an intervention mechanism can only be achieved through large-scale AB testing much akin to the (in)famous emotional contagion experiment (Kramer et al., 2014).

$\mathcal{CS}$ may not always have a positive effect. A study in Austria (n = 1285) reveals that countering hate speech increases polarization as attitudinal gaps increase between left-wing and right-wing participants (Schäfer et al., 2023).

The rich social science literature on studying the effectiveness of $\mathcal{CS}$ (Kim et al., 2023; Schäfer et al., 2023; Van Houtven et al., 2024) point to a palpable disconnect with the computational literature described in Section 4. Combining methodological advancements with stronger substantive analyses as observed in the social science literature will have a synergistic effect.

6 Research gaps and challenges

6.1 Coverage

Target groups. While $\mathcal{CS}$ research has touched upon several disadvantaged groups, much remains to be done in terms of coverage. For instance, Dalits are a historically marginalized group in India (Kapur et al., 2010). No $\mathcal{CS}$ research exists for Dalits in India. Extending $\mathcal{CS}$ resources to a diverse set of disadvantaged groups towards greater coverage thus merits sustained efforts.

Languages. $\mathcal{CS}$ resources span a small set of languages beyond English. Recent strategies to extend $\mathcal{CS}$ in low-resource languages include: (1) curating datasets in low-resource languages (e.g., (Nath et al., 2023; Hande et al., 2021)); (2) harnessing unsupervised methods to transfer resources to low-resource counterparts (KhudaBukhsh et al., 2020); and (3) leveraging generative AI advancements.

6.2 Resources

Datasets. Despite sustained efforts from several research groups, hate speech and $\mathcal{CS}$ have a stark resource mismatch. While there are hundreds of hate speech datasets, there exist only a handful of publicly available $\mathcal{CS}$ datasets. Several of these datasets are not curated from real-world social media conversations (e.g., (Chung et al., 2019)).

Workshops and shared tasks. $\mathcal{CS}$ research yet to receive similar importance as hate speech research in workshops hosted in premier AI and NLP conferences. Thus far, only one NLP workshop solely focused on $\mathcal{CS}$ has been conducted: The 1st Workshop on CounterSpeech for Online Abuse (CS4OA). A hope speech detection shared task is conducted as a part of the EACL workshop on Language Technology for Equality, Diversity, and Inclusion (LT-EDI) (Chakravarthi et al., 2022). Organizing more $\mathcal{CS}$ workshops around will stimulate further research in this area.

6.3 Annotation

Rater subjectivity and diversity. There is growing literature around annotation subjectivity that shows that human raters seldom have a broad consensus on what is offensive or hate speech (Sap et al., 2022; Weerasooriya et al., 2023). It is likely that with a large number of raters, $\mathcal{CS}$ datasets will also show inherent disagreement on what is $\mathcal{CS}$. Existing $\mathcal{CS}$ literature is yet to make a connection with the annotation subjectivity literature.

Participatory study design. Future $\mathcal{CS}$ datasets will benefit from annotation demographic transparency. While some of the datasets present detailed demographic information and are curated by expert NGO operators, a key question remains unanswered: *how well can a rater represent the values and opinions of the target group when the rater does not belong to the same identity group as the target group?* Simply put, can we rely on a straight person to effectively identify $\mathcal{CS}$ for transphobia, or won't it be better to ask a trans person to guide the AI system for $\mathcal{CS}$ strategies protecting trans people? Participatory AI frameworks (Khoramrouz et al., 2023) seek to build AI with active input from the stakeholders to improve system reliability. Introducing participatory AI and vicarious interactions (Weerasooriya et al., 2023) in dataset curation will improve the quality of $\mathcal{CS}$ datasets.

6.4 Counterspeech containing hate speech

Generative AI outputs are often unpredictable, marred with hallucinations, and other issues. A key concern echoed in several generative $\mathcal{CS}$ papers are *what if the generated $\mathcal{CS}$ are abusive or inappropriate?* (Mun et al., 2023; Gupta et al., 2023). The subsequent ethical conundrum is *how should humans label $\mathcal{CS}$ instances in the presence of abusive content?* This ethical dilemma was also discussed in the context of in-the-wild evaluation of *help speech* where annotators explained that given that the disenfranchised minorities were subjected to genocide, dehumanizing the perpetrators as animals was acceptable to them.

7 Best practices in $\mathcal{CS}$ research

7.1 Data collection and annotation

✔ *Use diverse and representative datasets*

☞ Collect data from multiple platforms to ensure a comprehensive analysis.

☞ Include multilingual and cross-cultural samples to study regional variations in $\mathcal{CS}$ effectiveness.

✔ *Ensure high-quality annotation*

☞ Use expert annotators or crowd-workers trained in hate speech nuances to label $\mathcal{CS}$ accurately.

☞ Implement inter-annotator agreement checks (e.g., Cohen's κ) to maintain consistency.

☞ Provide clear guidelines on distinguishing $\mathcal{CS}$ from general discussions.

✔*Balance hate speech and $\mathcal{CS}$ samples*
☞ Oversampling $\mathcal{CS}$ instances ensures that models do not become biased toward detecting only hate speech.
☞ Consider dataset augmentation techniques like paraphrasing & adversarial examples to improve model robustness.

7.2 Rigor in $\mathcal{CS}$ detection

✔*Use explainable AI for model interpretability*
☞ Avoid black-box models by leveraging attention mechanisms, feature attribution, & SHAP/LIME for interoperability.
☞ Provide justifications for how models differentiate counterspeech from hate speech.
✔*Benchmark against strong baselines*
☞ Compare new methods against traditional NLP techniques (e.g., SVM, logistic regression) and state-of-the-art models (BERT, GPT, T5, etc.).
☞ Use multiple evaluation metrics, including precision, recall, F1-score, and human evaluation.
✔*Account for context in detection models*
☞ Use thread-based classification instead of isolated sentence-level models, as $\mathcal{CS}$ effectiveness is often dependent on conversational flow.
☞ Train models on dialogue datasets where the interplay between hate speech and $\mathcal{CS}$ is preserved.

7.3 Best practices in $\mathcal{CS}$ generation

✔*Incorporate linguistic and psychological insights*
☞ $\mathcal{CS}$ should be polite, factual, intent-driven, and persuasive.
☞ Models should be fine-tuned on datasets labeled with $\mathcal{CS}$ strategies (e.g., humor, moral appeals, refutations).
✔*Use LLMs responsibly for $\mathcal{CS}$*
☞ Prompt engineering and reinforcement learning from human feedback (RLHF) can be used to align AI-generated $\mathcal{CS}$ with ethical norms.
☞ AI-generated $\mathcal{CS}$ should be evaluated through human studies before deployment.
✔*Address hate speech evolution*
☞ Continuous fine-tuning is necessary, as hate speech language evolves (e.g., use of coded language, dog whistles).
☞ Adversarial training can help models adapt to new forms of hate speech.
✔*Ensure ethical considerations in AI $\mathcal{CS}$*
☞ AI-generated responses should not accidentally reinforce harmful narratives.
☞ Transparency about AI-generated vs. human-written $\mathcal{CS}$ is essential for user trust.

7.4 Effective evaluation of $\mathcal{CS}$

✔*Conduct real-world impact studies*
☞ Use A/B testing on social media platforms to measure the actual impact of $\mathcal{CS}$ interventions.
☞ Analyze engagement metrics (likes, shares, replies) and hate speech reduction trends post-$\mathcal{CS}$ intervention.
✔*Measure psychological and behavioral impact*
☞ Assess whether $\mathcal{CS}$ reduces hate reinforcement or encourages non-hateful engagement from users.
☞ Consider conducting user surveys on the perceived effectiveness of different $\mathcal{CS}$ styles.
✔*Use ethical experimental designs*
☞ Ensure that $\mathcal{CS}$ interventions do not cause harm to victims or provoke hate speech escalation.
☞ Obtain informed consent when conducting human studies on $\mathcal{CS}$ engagement.

7.5 Cross-cultural and multilingual considerations

✔*Develop multilingual $\mathcal{CS}$ models*
☞ Training on low-resource languages and leveraging transfer learning ensures inclusivity.
☞ Consider cultural differences in what is considered offensive and how $\mathcal{CS}$ is perceived.
✔*Account for varying legal and social norms*
☞ Legal constraints on $\mathcal{CS}$ differ across regions (e.g., free speech laws in the U.S. vs. stricter hate speech laws in Germany).
☞ $\mathcal{CS}$ framing should align with local cultural norms to maximize effectiveness.

7.6 Future directions

✔*Incorporate multimodal $\mathcal{CS}$*
☞ Beyond text, $\mathcal{CS}$ research should explore image, video, and voice-based responses.
☞ Memes, GIFs, and visual $\mathcal{CS}$ strategies are gaining traction.
✔*Develop personalized $\mathcal{CS}$ strategies*
☞ AI models should adapt responses based on user psychology and engagement history.
☞ Personalized $\mathcal{CS}$ can be more persuasive than generic responses.
✔*Bridge the gap between academia and industry*
☞ Collaboration between researchers, social media platforms, and policymakers can lead to real-world deployment of $\mathcal{CS}$ models.
☞ Industry partnerships help in scaling up experimental models and integrating them into moderation systems.

Ethical statement

We survey existing methods in $\mathcal{CS}$ research and present our observations and recommendations. We do not present any novel dataset or methods in this paper.

Limitations

Although our survey covers several adjacent themes to $\mathcal{CS}$, such as *hope speech* and *help speech*, due to space limitations, we were unable to cover certain social science concepts that are related to $\mathcal{CS}$ (e.g., bending (Caponetto and Cepollaro, 2023)). Also, if space permitted, we would have liked to include a broader coverage of multilingual $\mathcal{CS}$.

References

Saharsh Agarwal, Uttara M Ananthakrishnan, and Catherine E Tucker. 2022. Deplatforming and the control of misinformation: Evidence from parler. *SSRN*.

M Ashida and M Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *WOAH*, pages 11–23.

B Bahador. 2021. Countering hate speech. In *The Routledge Companion to Media Disinformation and Populism*, pages 507–518. Routledge.

F Baider. 2023. Accountability issues, online covert hate speech, and the efficacy of counter-speech. *Pol. and Gov.*, 11(2):249–260.

David Bamman, Brendan O'Connor, and Noah Smith. 2012. Censorship and deletion practices in chinese social media. *First Monday*, 17(3).

Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. 2025. Safeinfer: Context adaptive decoding time safety alignment for large language models. In *AAAI*.

J Bartlett and A Krasodomski-Jones. 2016. Counter-speech on facebook. *Demos*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counter-speech on twitter: A field study. dangerous speech project.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024. Basque and spanish counter narrative generation: Data creation and evaluation. *Preprint*, arXiv:2403.09159.

Michael Bennie, Bushi Xiao, Chryseis Xinyi Liu, Demi Zhang, Jian Meng, and Alayo Tripp. 2025. Codeof-conduct at multilingual counterspeech generation: A context-aware model for robust counterspeech generation in low-resource languages. *Preprint*, arXiv:2501.00713.

Helena Bonaldi, Giuseppe Attanasio, Debora Nozza, and Marco Guerini. 2023. Weigh your own words: Improving hate speech counter narrative generation via attention regularization. In *CS4OA*, pages 13–28.

L Caponetto and B Cepollaro. 2023. Bending as counterspeech. *Eth. Theo. and Mor. Prac.*, 26(4):577–593.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, Subalalitha Cn, John Philip McCrae, Miguel Ángel García, Salud María Jiménez-Zafra, Rafael Valencia-García, Prasanna Kumaresan, Rahul Ponnusamy, and 1 others. 2022. Overview of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 378–388.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *CSCW*, 1:1–22.

Y-L Chung, Marco Guerini, and Rodrigo Agerri. 2021a. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132.

Y-L Chung, S. S Tekiroglu, and M. Guerini. 2020. Italian counter narrative generation to fight online hate speech.

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 *ACL*, pages 2819–2829.

Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of ACL-IJCNLP*, pages 899–914.

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. *Preprint*, arXiv:2402.07262.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *NAACL-HLT*, pages 2970–3005.

M Doğanç and I Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *CS4OA*, pages 1–12.

Brooke Erin Duffy and Colten Meisner. 2023. Platform governance at the margins: Social media creators' experiences with algorithmic (in) visibility. *Media, Culture & Society*, 45(2):285–304.

Arka Dutta, Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2024. Down the toxicity rabbit hole: A framework to bias audit large language models with key emphasis on racism, antisemitism, and misogyny. In *IJCAI*, pages 7242–50.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv:2107.08720*.

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 *WOAH*, pages 102–112.

Pierpaolo Goffredo, V Basile, B Cepollaro, and V Patti. 2022. Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts. In *WOAH*, pages 57–66.

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 *ACL*, pages 5792–5809.

S MD Halim, S M Halim, S Irtiza, Y Hu, Khan L, and B Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *IJCNN*.

Adeep Hande, Ruba Priyadharshini, Anbukkarasi Sampath, Kingston Pal Thamburaj, Prabakaran Chandran, and Bharathi Raja Chakravarthi. 2021. Hope speech detection in under-resourced kannada language. *CoRR*, abs/2108.04616.

D Hangartner, G Gennaro, S Alasiri, N Bahrich, A Bornhoft, J Boucher, B B Demirci, L Derksen, A Hall, M Jochum, M M Munoz, M Richter, F Vogel, S Wittwer, F Wüthrich, F Gilardi, and K Donnay. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *PNAS*, 118(50):e2116310118.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *ACL*, pages 3309–3326.

S Hassan and M Alikhani. 2023. Discgen: A framework for discourse-informed counterspeech generation. *ArXiv*, abs/2311.18147.

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. In *NAACL*, pages 6716–6733.

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

Manoel Horta Ribeiro, Homa Hosseinmardi, Robert West, and Duncan J Watts. 2023. Deplatforming did not decrease parler users' activity on fringe social media. *PNAS nexus*, 2(3):pgad035.

Shijia Huang, Jianqiao Zhao, Yanyang Li, and Liwei Wang. 2023. Learning preference model for LLMs via automatic preference data generation. In *EMNLP*, pages 9187–9199, Singapore.

Shuyu Jiang, Wenyi Tang, Xingshu Chen, Rui Tang, Haizhou Wang, and Wenxian Wang. 2023. Raucg: Retrieval-augmented unsupervised counter narrative generation for hate speech. *Preprint*, arXiv:2310.05650.

D Kapur, CB Prasad, L Pritchett, and DS Babu. 2010. Rethinking inequality: Dalits in uttar pradesh in the market reform era. *Eco. and Pol. Weekly*, pages 39–49.

A Kende and P Krekó. 2020. Xenophobia, prejudice, and right-wing populism in east-central europe. *Current Opinion in Behavioral Sciences*, 34:29–33.

Adel Khorramrouz, Sujan Dutta, and Ashiqur R. KhudaBukhsh. 2023. For Women, Life, Freedom: A Participatory AI-Based Social Web Analysis of a Watershed Moment in Iran's Gender Struggles. In *IJCAI*, pages 6013–6021.

Ashiqur R. KhudaBukhsh, Shriphani Palakodety, and Jaime G. Carbonell. 2020. Harnessing code switching to transcend the linguistic barrier. In *IJCAI*, pages 4366–4374.

Ashiqur R KhudaBukhsh, Rupak Sarkar, Mark S Kamlet, and Tom Mitchell. 2021. We don't speak the same language: Interpreting polarization through machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14893–14901.

Jae Yeon Kim, Jaeung Sim, and Daegon Cho. 2023. Identity and status: When counterspeech increases hate speech reporting and why. *Inf. Sys. Front.*, 25(5):1683–1694.

A DI Kramer, J E Guillory, and Hancock J T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS*, 111(24):8788.

Huije Lee, Young Ju NA, Hoyun Song, Jisu Shin, and Jong C. Park. 2022. ELF22: A context-based counter trolling dataset to combat Internet trolls. In *LREC*, pages 3530–3541.

Binny Mathew, Navish Kumar, Pawan Goyal, and Animesh Mukherjee. 2020. Interaction dynamics between hate and counter users on twitter. In *CoDS-COMAD*, pages 116–124. ACM.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *ICWSM*, volume 13, pages 369–380.

Karsten Müller and Carlo Schwarz. 2021. Fanning the flames of hate: Social media and hate crime. *J. of the Eur. Econ. Asso.*, 19(4):2131–2167.

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. In *Findings of EMNLP*, pages 9759–9777.

Jimin Mun, Cathy Buerger, Jenny T Liang, Joshua Garland, and Maarten Sap. 2024. Counterspeakers' perspectives: Unveiling barriers and ai needs in the fight against online hate. In *CHI*, CHI '24.

Tanusree Nath, Vivek Kumar Singh, and Vedika Gupta. 2023. Bonghope: An annotated corpus for bengali hope speech detection.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models, may 2023. *URL http://arxiv. org/abs/2305.12827*.

Aswini Kumar Padhi, Anil Bandhakavi, and Tanmoy Chakraborty. 2025. Counterspeech the ultimate shield! multi-conditioned counterspeech generation through attributed prefix learning. *Preprint*, arXiv:2505.11958. To Appear ACL 2025.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020a. Hope speech detection: A computational analysis of the voice of peace. In *ECAI 2020*, pages 1881–1889. IOS Press.

Shriphani Palakodety, Ashiqur R KhudaBukhsh, and Jaime G Carbonell. 2020b. Voice for the voiceless: Active sampling to detect comments supporting the rohingyas. In *AAAI*, pages 454–462.

Kaike Ping, Anisha Kumar, Xiaohan Ding, and Eugenia Rho. 2024. Behind the counter: Exploring the motivations and barriers of online counterspeech writing. *Preprint*, arXiv:2403.17116.

E Pronoza, Polina Panicheva, Olessia Koltsova, and Paolo Rosso. 2021. Detecting ethnicity-targeted hate speech in russian social media texts. *Inf. Proc. & Mgmt.*, 58(6):102674.

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 *EMNLP-IJCNLP*, pages 4755–4764.

Punyajoy Saha, Mithun Das, Binny Mathew, and Animesh Mukherjee. 2023. Hate speech: Detection, mitigation and beyond. In *WSDM*, pages 1232–1235.

Punyajoy Saha, Abhilash Datta, Abhik Jana, and Animesh Mukherjee. 2024. Crowdcounter: A benchmark type-specific multi-target counterspeech dataset. In *CoNLL*.

Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022. CounterGeDi: A Controllable Approach to Generate Polite, Detoxified and Emotional Counterspeech. In *IJCAI*, pages 5157–5163.

S Saha and R Srihari. 2024. Consolidating strategies for countering hate speech using persuasive dialogues. *Preprint*, arXiv:2401.07810.

E M Saltman and J Russell. 2014. White paper–the role of prevent in countering online extremism. *Quilliam publication*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *NAACL-HLT*, pages 5884–5906.

S Schäfer, I Rebasso, M M Boyer, and A M Planitzer. 2023. Can we counteract hate? effects of online hate speech and counter speech on the perception of social groups. *Communication Research*, page 00936502231201091.

N Strossen. 2018. *Hate: Why we should resist it with free speech, not censorship*. OUP.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of ACL*, pages 3099–3114.

H Tuck and T Silverman. 2016. The counter-narrative handbook. *Institute for Strategic Dialogue*, 1.

A Tyagi, Anjalie Field, Priyank Lathwal, Yulia Tsvetkov, and Kathleen M. Carley. 2020. A computational analysis of polarization on indian and pakistani social media. In *SocInfo*, volume 12467 of *LNCS*, pages 364–379.

E Van Houtven, S B Acquah, M Obermaier, M Saleem, and D. Schmuck. 2024. 'you got my back?'severity and counter-speech in online hate speech toward minority groups. *Media Psychology*, pages 1–32.

Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. 2020. Detecting East Asian prejudice on social media. In *WOAH*, pages 162–172.

Haiyang Wang, Zhiliang Tian, Xin Song, Yue Zhang, Yuchen Pan, Hongkui Tu, Minlie Huang, and Bin Zhou. 2024. Intent-aware and hate-mitigating counterspeech generation via dual-discriminator guided llms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language*

Resources and Evaluation (LREC-COLING 2024), pages 9131–9142.

Tharindu Cyril Weerasooriya, Sujan Dutta, Tharindu Ranasinghe, Marcos Zampieri, Christopher M. Homan, and Ashiqur R. KhudaBukhsh. 2023. Vicarious offense and noise audit of offensive speech classifiers: Unifying human and machine disagreement on what is offensive. In *EMNLP*, pages 11648–11668.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on Twitter. In *WOAH*, pages 57–62.

Clay H. Yoo, Shriphani Palakodety, Rupak Sarkar, and Ashiqur R. KhudaBukhsh. 2021. Empathy and hope: Resource transfer to model inter-country social media dynamics. In *Workshop on NLP for Positive Impact*, pages 125–134.

Xinchen Yu, Eduardo Blanco, and Lingzi Hong. Hate speech and counter speech detection: Conversational context does matter.

Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on tiktok. *Policy & Internet*, 14(1):79–95.

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