Annotation Guidelines for Evaluating Effectiveness of Counter Speech

1 Introduction

This document describes the guidelines to annotate the effectiveness of counter-narratives (CNs) in the field of hate speech (HS).
In this work, we define a counter-narrative as:

a non-aggressive response that offers feedback through fact-bound arguments and is considered as the most effective approach to withstand hate messages ([4], [12]).

In this definition, it appears the word *effective*, but how to define and measure it remains an open question. This work aims at finding a suitable definition of effectiveness and a comprehensive way to measure it.
The generation and evaluation of counter-narratives attacking hate speech is a relatively new field of research in the domain of Natural Language Processing (NLP) tasks. The majority of studies on counter-narratives address different methods for creating CN datasets [8, 10, 6, 5], and tackle different counter-narrative generation methodologies [2, 13, 9, 11, 16, 14].
This work is one of the first explicitly focusing on evaluating the effectiveness of CNs. In fact, few studies have been conducted on this topic [2, 11]. Evaluation of generated text, and specifically of counter-narratives, remains a difficult task. Creating effective responses is challenging due to the wide range of acceptable outputs and the difficulty in defining what constitutes a good response. This complicates the design of evaluation schemes. Some constrained generation tasks, like machine translation, have established evaluation metrics such as "adequacy" and "fluency," but many other generation tasks lack standardized evaluation criteria. For counter-narratives, there are no established evaluation methods. Previous research has suggested various human evaluation metrics, including suitableness, informativeness, intra-coherence [7], diversity, relevance, language quality [16], offensiveness, and stance [3]. The aim of this study is to develop a comprehensive evaluation framework for assessing the effectiveness of counter speech. Building on previous related work, the framework combines six key components: Clarity, Evidence, Audience Adaptation, Emotional Appeal, Rebuttal, and Fairness. This framework will be used to enrich existing counter speech datasets by incorporating these new evaluation dimensions, enabling a multi-faceted analysis of counter speech. Additionally, we aim to build a classifier that can identify effective counter speech by providing an overall assessment score while highlighting specific areas for improvement.
Currently, there is no universally agreed-upon definition of what constitutes an effective counter speech. Existing research often evaluates the quality of counter speech using effectiveness as one of the metrics. For instance, [15] define counter speech as effective if it helps targets feel better and fosters empathy in bystanders toward the targets.
In this study, we define an effective counter-narrative, particularly in the context of countering hate speech, as a strategically designed message aimed at challenging or providing an alternative to hateful content. The primary goal of such counter speech is to address and transform the underlying beliefs and attitudes that incite harmful ideas. This approach emphasizes the importance of understanding the deeper reasons behind negative beliefs with the aim of obtaining changes in attitudes, beliefs, or behaviors related to hateful ideologies.
Therefore, in order to be effective, a counter speech should satisfy two conditions, i.e. it has to be:

1. **challenging** a harmful hate speech;
2. **persuading** and changing the ideas and behaviors.

2 Datasets description

- **CONAN**. This dataset [6] is a multilingual and expert-based dataset of hate speech/counter-narrative pairs in English, French and Italian, focused on Islamophobia. The dataset consists of 4078 pairs over the 3 languages. The dataset is augmented through translation (from Italian/French to English) and paraphrasing, which brought the total number of pairs to 14.988. We filtered out messages that are not in English.
- **Twitter Dataset**. This dataset [1] contains 5,652 hateful tweets and replies. The authors analyzed their fine-grained relationships by indicating whether the reply (a) is hate or counter hate speech, (b) provides a justification, (c) attacks the author of the tweet, and (d) adds additional hate. We have automatically extracted the target of hate, which was not present in the original dataset, and we checked manually that the prediction was correct. (This dataset is publicly available only with twitter IDs, for privacy issues. We asked the authors to provide it with all the fields, and they kindly agree).

These datasets have been chosen in order to analyze different types of counter-narratives: some of them have been accurately written by trained NGO operators, while others have been collected from Twitter, written by non trained online social media users. In this way, we aim at analyzing the difference in effectiveness between diverse types of counter-narratives' authors (expert vs non-expert).

3 Evaluation metrics

In this section, we list the effectiveness components, together with their definition, scoring range, and examples.

1. **Clarity**

   Clarity in an effective counter speech means being easily understood and straightforward, ensuring the message is unambiguous, accessible, and relatable to the audience. This means that effective arguments should be clear and logically sound. In this way, it is easier for the audience to follow the reasoning and understand the main points.

   The evaluation of Clarity is addressed using the following scoring:

   3: The counter-narrative is clearly articulated, logically structured, and directly addresses the specific topic of the hateful message. Additionally, the counter-narrative is free from irrelevant information, unnecessary complexity, or ambiguity.

   2: The counter-narrative is mostly clear and well structured, but it is very general, and could be applied to multiple hateful messages.

   1: The counter-narrative is not clear and has a structure that is not understandable.

   Following there are examples of counter speech with high and low Clarity scores, together with an explanation.

   Example with high Clarity (3 points):

   HS: *People with disabilities are a burden on society and can't contribute meaningfully.*

   CS: *People with disabilities contribute to society in many ways, both economically and socially. Many hold jobs, run businesses, and actively participate in their communities. When given proper support, people with disabilities thrive in education and employment, making meaningful contributions like anyone else.*

   Explanation: The language is simple and direct, and there is no ambiguity. It has a logical structure, starting with a counter to the "burden" claim and then providing a clear explanation of how people with disabilities contribute to society. It is specific to the topic, as it directly

addresses the claim that people with disabilities "are a burden" and provides a clear counter to that idea. Finally, the counter-narrative remains on-topic and avoids unnecessary details or complexity.

Example with low Clarity (1 point):

HS: *People with disabilities are a burden on society and can't contribute meaningfully.*

CS: *Well, not everyone is the same, and it's complicated. Disabilities are different, like, some people have disabilities that don't even stop them from doing stuff. And society has a lot of issues anyway, so saying someone is a burden is just wrong.*

Explanation: The counter-narrative relies on phrases like "it's complicated" without providing any specific information or arguments: this is vague. Moreover, it strays far from the original topic, introducing broad and unrelated concepts like "society's issues" and "life isn't just about work." Finally, it has a poor structure, because there's no logical flow between the points, and it's hard to understand what the counter-hater is trying to argue.

2. **Evidence**

Evidence in a counter speech is credible, relevant information that supports claims and reinforces validity, showing the argument is based on facts, research, or logic rather than opinion. Effective arguments are supported by relevant evidence and examples that support the argument's claims and make them more compelling to the audience. They should be specific to the topics addressed in the hateful message.

The evaluation of Evidence is addressed using the following scoring:

- **3**: the counter-narrative is highly informative, mentioning multiple pieces of information (e.g., specific data, an event, or a person), as supporting evidence, that are very specific to what is present in the hate speech text. They are all sound and acceptable.
- **2**: the counter-narrative mentions only one piece of information, as supporting evidence, that is specific to the content of the hateful message.
- **1**: the counter-narrative does not present any additional information, that can be used as supporting evidence.

Following there are examples of counter speech with high and low Evidence scores, together with an explanation.

Example with high Evidence (3 points):

HS: *Jews control the banks, media, Hollywood, even the u.s. government.*

CS: *This claim is just a byproduct of "The Protocols of the Elders of Zion"—a late 19th-century forgery about a supposed global Jewish conspiracy, produced by members of the czar's secret police and exposed as a fiction by the Times of London. Jews were accused of desecrating the sacramental bread used in communion, poisoning wells and spreading the plague. Nowadays they are accused of controlling Hollywood and the banks.*

Explanation: The counter-narrative provides several additional information that are not present in the HS: it describes in detail the origin of the conspiracy theory on which the HS is based, mentioning (1) the name of the false document ("The Protocols of the Elders of Zion") (2) when it was created (late 19th-century) (3) who produced it (the members of the czar's secret police) (4) who exposed it (the Times of London). All the information are correct.

Example with low Evidence (1 point):

HS: *I am sure California is full of sick self hating whites , who think that for the good of all the black brown and yellow peoples of the world , whites need to commit racial suicide and disappear from their face of the earth for their sake and won't that be just dandy!*

CS: *Promoting harm or extinction of any race is not the solution to societal issues. We should strive for unity, understanding, and respect among all races, as diversity and coexistence are what truly benefit humanity.*

Explanation: The counter-narrative just generally claims that promoting harm is not a solution, but does not provide any additional information or facts supporting this claim.

3. **Emotional Appeal**

Emotional appeal is a persuasive technique that influences an audience by evoking emotions like empathy, fear, or anger. In counter-speech, it uses language, stories, or imagery to make arguments more compelling and personally impactful.

The evaluation of Emotional Appeal is addressed using the following scoring:

1: the counter-narrative makes use of words or expressions that evoke emotions - either positive or negative. Sarcasm and irony are included in this category.

0: the counter-narrative does not make use of words or expressions that evoke emotions.

Following there are examples of counter speech with high and low Emotional Appeal scores, together with an explanation.

Example of a counter-narrative making use of emotional appeal (1 point):

HS: *Women getting into the labour market has caused the downfall of Western civilisation, they should be at home raising children.*

CS: *Imagine a world without the contributions of women in the workforce—no nurses to care for us, no teachers to educate our children, no scientists, doctors, or leaders driving innovation and progress. Women are the backbone of both our homes and our communities. Their strength and dedication, whether in raising children or in professional roles, uplift and enrich our society. Dismissing their invaluable contributions and potential is not only unjust but also a disservice to the future we all hope to build—a future of equality, compassion, and shared success.*

Explanation: The counter speech aims to resonate with the reader on a deeper level, encouraging empathy and a reevaluation of the initial hateful message. This is achieved by using different techniques. The use of imagery by asking the reader to envision a world without women's contributions in various vital roles, it evokes a sense of loss. The CN emphasizes shared societal values like equality, compassion, and progress appeals to the reader's sense of justice. Finally, the CN, by referencing the dual role of women in both family (raising children) and community (professional contributions), builds an emotional connection, emphasizing that women's impact is both personal and communal.

Another example of a counter-narrative making use of emotional appeal (1):

HS: *Women getting into the labour has caused the downfall of Western civilization, they should be at home raising children.*

CS: *Absolutely! Because, as we all know, the world was completely perfect before women started working. No wars, no economic crises, no social problems—just smooth sailing all the way!*

Explanation: The counter speech employs sarcasm to exaggerate the idea that Western civilization was perfect before women entered the workforce, exposing the absurdity of the original claim. By ironically downplaying the contributions of women in various fields, the response highlights how critical those contributions actually are.

Example of a counter speech making no use of emotional appeal (0 points):

HS: *Women getting into the labour market has caused the downfall of Western civilisation, they should be at home raising children.*

CS: *The idea that women entering the labor market has caused the downfall of Western civilization is not supported by evidence. Numerous studies have shown that women's participation in the workforce boosts economic growth, enhances innovation, and improves overall productivity. Furthermore, modern societies benefit from diverse perspectives, and excluding half the population from contributing professionally would be economically and socially detrimental. Gender equality in the labor market is a hallmark of progressive and thriving societies, not their downfall.*

Explanation: This counter speech avoids emotional appeal and instead relies on logical reasoning and factual statements for the following reasons. It uses evidence-based statements, as the reference to "numerous studies". This approach relies on empirical evidence rather than appealing to the reader's emotions. It also uses logical reasoning to explain why excluding women from the workforce would be detrimental to society. Lastly, the counter-narrative uses a neutral and objective language, avoiding emotionally charged words or phrases.

4. **Rebuttal**

A rebuttal is a response that directly addresses and counters an opposing argument or claim. It aims to weaken the opposing viewpoint by presenting evidence, logical reasoning, or highlighting flaws and inconsistencies in the original argument. A strong rebuttal not only defends one's position but also demonstrates why the opposing perspective is less valid or less persuasive. Rebuttal is used to assess how much the arguments of the counter speech are effective to anticipate and address potential counterarguments or objections from the audience.

The evaluation of Rebuttal is addressed using the following scoring :

- **3**: The counter-narrative provides multiple rebuttals (two or more), each of which clearly responds to a specific part of the original claim. Rebuttals use evidence, logic, or reasoning to challenge or disprove the hateful message effectively. These rebuttals directly target different aspects of the claim, leaving little room for the original argument to stand.
- **2**: The counter-narrative presents only one rebuttal, which responds to a central point of the original claim. This rebuttal provides some evidence or reasoning, but focuses on addressing a single issue or argument within the hateful message.
- **1**: The counter-narrative does not directly counter any part of the hateful message. It may discuss related topics or make general statements but lacks a clear rebuttal that challenges the original claim with evidence or reasoning.

Following there are examples of counter speech with high and low Rebuttal scores, together with an explanation.

Example with a high score for rebuttal (3 points):

HS: *I am done with Islam and ISIS. All Muslims should be sent to their homeland. Britain will be better without their violence and ideology.*

CS: *Saying "All Muslims should be sent to their homeland" unfairly blames all Muslims for the actions of ISIS. First, it's crucial to understand that the vast majority of Muslims are peaceful and have no connection to extremism. Associating an entire religion with the actions of a few extremists is both incorrect and unjust. Second, Muslims in Britain make significant contributions in many areas such as healthcare, education, and business. Removing them would not make Britain better but would harm the country's diversity and strength.*

Explanation: The counter-argument presents two rebuttals. The first one is addressing the unfair generalization: saying 'All Muslims should be sent to their homeland' unfairly blames all

Muslims for the actions of ISIS. This rebuttal addresses the logical fallacy of generalization. It clarifies that blaming all Muslims for the actions of ISIS is an unfair stereotype.

The second rebuttal is about ignoring Muslims' contributions to society. This rebuttal focuses on the positive contributions of Muslims to British society. It counters the idea that Britain would be better without Muslims by highlighting their roles in critical sectors like healthcare, education, and business.

Example with a low score for rebuttal (1 point):

HS: *I am done with Islam and ISIS. All Muslims should be sent to their homeland. Britain will be better without their violence and ideology.*

CS: *It's wrong to blame all Muslims for the actions of a few. We should focus on unity and understanding, not division.*

Explanation: This counter-argument does not include rebuttals because it does not directly address or refute specific points made in the original hateful message. Instead, it presents a general positive perspective on the contributions and inclusion of Muslims.

5. **Audience Adaptation**

Audience adaptation in a counter speech means tailoring the message to the audience's values, knowledge, and preferences to make it more persuasive. This involves using the right language, tone, and examples to connect with them effectively.

The evaluation of Audience Adaptation is addressed using the following scoring:

1: The counter-narrative uses language that matches the level of complexity or simplicity in the original hateful message. Tone and vocabulary are adapted to reflect the same style and level of understanding as the hate speech, ensuring the counter-narrative resonates with the audience on a similar linguistic level.

0: The counter-narrative uses language that is either more complex or more simplistic compared to the original hateful message. The tone or vocabulary of the counter-narrative does not align with the audience's expected language, potentially making it harder for the message to resonate or connect.

Following there are examples of counter speech with high and low Audience Adaptation scores, together with an explanation.

Example with high audience adaptation (1 point):

Example (Simple language):

HS: *Immigrants are lazy and just take our jobs without doing any real work.*

CN: *That's just not true. Immigrants work hard, often in jobs others don't want, and they help keep the economy running.*

Example (Complex language):

HS: *The influx of immigrants has degraded the labor market, contributing nothing to societal advancement.*

CN: *On the contrary, immigrants contribute significantly to the labor market, bringing diversity and skills that enrich society and drive innovation.*

Example with low audience adaptation (0 point):

Example (Different language level – more complex):

HS: *Immigrants are lazy and just take our jobs without doing any real work.*

CN: *That's a gross mischaracterization of the labor contributions of immigrants. Empirical data shows that immigrant workers bolster economic growth by filling critical gaps in the workforce.*

Example (Different language level – more simplistic):

HS: *The influx of immigrants has degraded the labor market, contributing nothing to societal advancement.*

CN: *Immigrants help us, they're not bad. They do important jobs and help our country.*

6. **Fairness**

Fairness in a counter-argument means presenting opposing views accurately, acknowledging valid points, and avoiding distortion or bias. It involves using a respectful tone and offering a balanced, thoughtful response to enhance credibility, and avoiding attacks or insults. Effective counter-narratives use appropriate language, by respecting freedom of expression, without censoring the opposing viewpoint.

The evaluation of Fairness is addressed using the following scoring:

3: The counter-narrative is polite, it does not attack the hater and it does not use any swear-word.

2: The counter-narrative is mostly polite, and does not attack the hater. However, it uses a swearword, or its content can be interpreted as slightly offensive.

1: The counter-narrative has a hateful content, uses a swearword, or it attacks the hater.

Following there are examples of counter speech with high and low Fairness scores, together with an explanation.

Example with high Fairness (3 points):

HS: *I am sure California is full of sick self hating whites , who think that for the good of all the black brown and yellow peoples of the world, whites need to commit racial suicide and disappear from their face of the earth for their sake and won't that be just dandy!*

CS: *Promoting harm or extinction of any race is not the solution to societal issues. We should strive for unity, understanding, and respect among all races, as diversity and coexistence are what truly benefit humanity.*

Explanation: The counter-narrative is polite, it attacks the message and not the hater, and it does not use violent language.

Example with low Fairness (1 points):

HS: *They should make all know about this book his friends as that they know like are the things , we need that all the whites know like they are the things and the Jewish threat and negroid on our race.*

CS: *The negroid race is not threatening, it is a good thing to have in our societies.*

Explanation: The meaning of the counter-narrative is not hateful, but it uses a violent language.

Example with low Fairness (1 point):

HS: *When I was last time in South Africa , we went to Sun City and saw little black monkey children swimming , funniest thing was watching drying water off their faces , it looks exactly like baboon or something , way they do it , whites have certain mannerisms with doing things, they have completely ape-like ones.*

CS: *The main difference is that apes have brains. Don't you think?*

Explanation: The counter-narrative has a hateful content.

References

[1] Abdullah Albanyan and Eduardo Blanco. Pinpointing fine-grained relationships between hateful tweets and replies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10418–10426, June 2022. doi: 10.1609/aaai.v36i10.21284.

[2] Mana Ashida and Mamoru Komachi. Towards automatic generation of messages countering online hate speech and microaggressions. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat, editors, *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.2.

[3] Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.397.

[4] Susan Benesch. Countering dangerous speech: New ideas for genocide prevention, 2014.

[5] Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroglu, and Marco Guerini. 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. Association for Computational Linguistics, December 2022.

[6] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 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, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1271.

[7] Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.79.

[8] Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.250.

[9] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.318.

[10] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. Thou shalt not hate: Countering online hate speech. In *Thirteenth International AAAI Conference on Web and Social Media*, 2019.

[11] Jimin Mun, Emily Allaway, Akhila Yerukola, Laura Vianna, Sarah-Jane Leslie, and Maarten Sap. Beyond denouncing hate: Strategies for countering implied biases and stereotypes in language. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9759–9777, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.653.

[12] Carla Schieb and Mike Preuss. Governing hate speech by means of counterspeech on facebook. In *66th ICA Annual Conference*, pages 1–23, Fukuoka, Japan, 2016.

[13] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. Generating counter narratives against online hate speech: Data and strategies. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.110.

[14] Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*, 2022.

[15] Yi Zheng, Björn Ross, and Walid Magdy. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In Yi-Ling Chung, Helena Bonaldi, Gavin Abercrombie, and Marco Guerini, editors, *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia, September 2023. Association for Computational Linguistics.

[16] Wanzheng Zhu and Suma Bhat. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.12.