From Generic to Personalized: Investigating Strategies for Generating Targeted Counter Narratives against Hate Speech

Mekselina Doğanç
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
`m.doganc@student.vu.nl`

Ilia Markov
CLTL, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
`i.markov@vu.nl`

Abstract

The spread of hate speech (HS) in the digital age poses significant challenges, with online platforms becoming breeding grounds for harmful content. While many natural language processing (NLP) studies have focused on identifying hate speech, few have explored the generation of counter narratives (CNs) as means to combat it. Previous studies have shown that computational models often generate CNs that are dull and generic, and therefore do not resonate with hate speech authors. In this paper, we explore the personalization capabilities of computational models for generating more targeted and engaging CNs. This paper investigates various strategies for incorporating author profiling information into GPT-2 and GPT-3.5 models to enhance the personalization of CNs to combat online hate speech. We investigate the effectiveness of incorporating author profiling aspects, more specifically the age and gender information of HS authors, in tailoring CNs specifically targeted at HS spreaders. We discuss the challenges, opportunities, and future directions for incorporating user profiling information into CN interventions.

Content Warning: *This paper discusses examples of hateful content. The authors do not support the use of hateful language, nor any of the hateful representations quoted below.*

1 Introduction

Hate speech (HS) is spreading through various online channels, targeting vulnerable individuals and minority groups, and leading to discrimination and division. The spread of HS can be motivated by fear, ignorance, and prejudice, with severe consequences such as increased suicide rates (O'Keeffe et al., 2011). Since there is a fine line between free speech and hate speech, monitoring the removal or blocking of online HS is challenging and may lead to over-blocking (Gagliardone et al., 2015; Markov and Daelemans, 2021). Instead of enforcing silence through blocking, "more speech" can be used, e.g., through the generation of counter narratives (CNs), which promotes understanding and different perspectives. Benesch (2014) defines counter narratives as "non-negative responses to hate speech, targeting and counteracting extreme statements through fact-bound arguments or alternative perspectives". Counter-narrative generation plays a crucial role in educating and enlightening those who may be influenced by hate speech or exposed to biased narratives (Pariser, 2011). Providing well-articulated and evidence-based responses, counter narratives can expose the fallacies and misinformation propagated by hate speech, enabling individuals to make informed judgments and decisions. NGOs work for exactly this purpose by training operators to write counter narratives[1], however, there is a great imbalance between the number of hate speech shared every day and the responses that can be written by NGO operators. To tackle this issue, natural language generation (NLG) studies have suggested automatic CN generation (Qian et al., 2019; Tekiroğlu et al., 2020a; Chung et al., 2021; Tekiroğlu et al., 2022). On top of that, profiling hate speech authors has gained attention in natural language processing (NLP) and sociolinguistic communities, providing valuable insights into HS authors' motivations and enhancing hate speech detection systems by incorporating author profiling information (Schmidt and Wiegand, 2017; Rangel et al., 2021; Monti et al., 2022; Hilte et al., 2023). To the best of our knowledge, exploring author profiling aspects within personalized counter-narrative interventions remains an unexplored area of research despite its potential to effectively counter

---

[1] https://getthetrollsout.org/stoppinghate

Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pages 1–12
September 11, 2023 ©2023 Association for Computational Linguistics

hate speech online. Consider the following example with two possible CN responses to hateful content[2]:

(1) HS: Women must not have the right to choose.

CN1: Women should have equal rights in a free society, without discrimination.

CN2: It can be difficult to adjust to a changing society, but it's important to recognize that gender equality and freedom are not just ideals; they are human rights that should be respected and upheld at all times. Women should have the same rights and opportunities as men, without any discrimination or prejudice. Older generations can help foster a culture of equality and respect by acknowledging that everyone deserves to be seen and treated equally, regardless of gender or any other background.

CN2 is more personalized and likely to be more engaging for a HS author than CN1. We explore ways how to make counter-narrative interventions more personalized and effective in withstanding the propagation of online hate speech. Our approach is based on replicating the study by Tekiroğlu et al. (2022) with GPT-2 (Radford et al., 2019) and exploring different strategies for injecting author profiling aspects into the model with the aim of generating more targeted counter narratives. Additionally, we experiment with GPT-3.5[3] to investigate its potential for personalizing counter narrative based on profiling characteristics. Findings suggest that while GPT-2 struggles with using the injected author profiling aspects to generate more personalized CNs, GPT-3.5 has the ability to incorporate profiling information, such as authors' age and gender, to generate more personalized and engaging responses against hate speech. Ultimately, this paper emphasizes the significance of effective counter narratives as a valuable tool for fostering a more inclusive and harmonious digital environment, free from the damaging effects of hate speech.

2 Related Work

2.1 Counter-narrative generation

Several studies have delved into the domain of CN generation, exploring different approaches and techniques to address the challenge of generating effective and persuasive responses to online hate speech (Qian et al., 2019; Tekiroğlu et al., 2020b; Chung et al., 2021; Tekiroğlu et al., 2022).

Studies that focused on CN generation report that CN interventions remain dull and generic. For instance, Qian et al. (2019) used different computational models (Seq2Seq, VAE and RL) to generate a response to hate speech that can mitigate its use during a conversation, reporting that the responses are largely commonplace and sometimes irrelevant. Chung et al. (2021) proposed a knowledge-grounded counter-narrative generation architecture that combats hallucination phenomena by incorporating factual information from external knowledge sources, enabling the model to generate informed and reliable counter narratives. However, even with a more knowledge-bound approach, the generation of counter narratives still lacks personalization and engagement with an individual hate speech author. Furthermore, the scarcity of comprehensive and representative datasets poses an additional challenge in the field of counter-narrative generation with only few available datasets that contain HS-CN pairs in the English language (Qian et al., 2019; Chung et al., 2019; Tekiroğlu et al., 2020a; Fanton et al., 2021).

2.2 Author profiling

A number of studies have focused on author profiling and its benefits, for instance, to gain insights into the underlying motivations, biases, and intentions of HS authors (Casavantes et al., 2023; Ebrahimi and Dou, 2016; Johannsen et al., 2015; Li and Tuzhilin, 2019; Zeng et al., 2019; Rangel et al., 2021). Schmidt and Wiegand (2017) provide a survey on traditional approaches to hate speech detection, which have predominantly relied on lexical and syntactic features, focusing on the content of the messages themselves. While such methods have provided valuable insights, they often fall short when confronted with the subtleties and nuances present in hate speech instances. It is in this context that the concept of author profiling emerges as a compelling avenue for deeper analysis.

For instance, the study by Casavantes et al. (2023) focused on abusive language detection lever-

---

[2] HS-CN1 pair extracted from the dataset we work with (Fanton et al., 2021), while CN2 is generated by one of the strategies described in this paper (GPT-3.5).

[3] https://openai.com/

aging posts' and authors' metadata. The study aimed to enhance the identification of abusive content by considering contextual information associated with individual posts and authors. The authors enriched textual features from posts with author-level metadata (i.e., information about the author, such as their account age, follower count, and verified status) significantly improving the detection accuracy of abusive comments across various categories.

A conceptually similar work to ours is by Zeng et al. (2019) whose goal was the development of a system that automatically generates personalized comments based on user profiles. Their objective was to enhance user engagement and interaction by tailoring comments to individuals' preferences and characteristics. The study presented an approach that leveraged user profile information, such as demographic data, interests, and preferences, to generate comments that align with the user's profile. The authors collected and analyzed user profiles to extract relevant features, such as age, gender, location, and stated interests, and trained a machine learning model (LSTM and Seq2Seq) using a dataset of existing comments and associated user profiles. This model learned to map user profiles to comment content, capturing the relationship between user characteristics and the language used in comments. The trained model was then used to generate personalized comments for new users based on their profile information.

As these and several other studies suggest, profiling hateful content creators can improve hate speech detection systems by incorporating profiling and stylometric information (Rangel et al., 2021; Markov et al., 2021; Casavantes et al., 2023), shed light on the HS phenomenon itself, and help to gain deeper insights into the reasons behind posting hateful content online (Rangel et al., 2021; Monti et al., 2022). To the best of our knowledge, profiling hate speech spreaders has not been explored in the context of personalization of counter-narrative interventions. To bridge this gap, we leverage HS authors' age and gender as profiling aspects to explore the effect it has on making CNs more personalized with respect to HS authors.

3 Data

Fanton et al. (2021) released a high-quality and high-quantity dataset, `MultiCONAN`, which is necessary for the generation of effective counter narratives, specifically due to the topic's sensitivity. Following Tekiroğlu et al. (2022), we use this dataset to explore various strategies aimed to generate more personalized counter narratives. The quality is assured through niche-sourcing—a niche group of experts in data collection, i.e., NGO operators who have the experience in generating counter narratives that will hinder the beginning of a negative spiral. `MultiCONAN` consists of HS-CN pairs for English comprising several targets of hate speech. The data was created with a human-in-the-loop approach (human input and expertise are incorporated into the process of gathering or labeling the data) (Fanton et al., 2021).

The dataset consists of 5,003 HS-CN pairs covering multiple targets presented in Table 1.

Target	# of HS-CN pairs
Muslims	1,335
Migrants	957
Women	662
LGBT+	650
Jews	598
POC	364
Disabled	220
Other	217

Table 1: The number of HS-CN pairs per target of HS.

Since this dataset does not include profiling information about the hate speech authors, we used the TEXTGAIN text analytics API[4] to estimate the authors' age (younger/older than 25) and gender (male/female). The age classification of TEXTGAIN determines whether a text is written by an adolescent or an adult leveraging, for instance, stylometric choices such as the usage of informal language, and the repeated occurrence of certain topical features. In a similar way, the gender prediction estimates whether a text is written by a man or a woman using lexical (e.g., usage of determiners, personal pronouns) and structural features (e.g., sentence structure)[5]. Hence, profiling aspects encompass binary information about age and gender—the author is either male or female and younger or older than 25 years. We incorporate this information into the `MultiCONAN` dataset using various strategies to explore the effect it has on CN personalization.

---

[4] https://devops.textgain.com/

[5] https://devops.textgain.com/#tag/Profiling

Since we are interested in identifying how to best generate personalized CNs, we manually revised the HS-CN pairs to filter out those HS instances for which age and gender did not seem to bear importance to the personalization of the counter narratives. For instance, the following HS instance: “Physically disabled folks are all angry and useless” is hard to personalize based solely on information about the age and gender of the author. After revision, the dataset was reduced to a total of 913 HS-CN pairs with age and gender information about HS authors. This subset is used for the strategies aimed to leverage author profiling information described in the following section.

4 Methods

The present research replicates the approach presented in the study by Tekiroğlu et al. (2022). We used the GPT-2 model (Radford et al., 2019) that showed promising results with the hyperparameters and decoding mechanisms reported in the continuation of this section. We designed a suite of experiments to explore the effect of incorporating profiling information: using GPT-2 to generate counter narratives while utilizing profiling information exclusively in the test set, fine-tuning GPT-2 with profiling information, fine-tuning GPT-2 with more detailed profiling information, generating personalized CNs using GPT-3.5 and fine-tuning GPT-2 with those personalized CNs. In the continuation, we provide more details into strategies used to incorporate author profiling characteristics.

4.1 Replication

We replicated the approach presented by Tekiroğlu et al. (2022) with the GPT-2 model (Radford et al., 2019), hence the first step was fine-tuning the model. We used the same ratio as in Tekiroğlu et al. (2022) and split the dataset into 80-10-10 for training, validation, and testing, respectively. A hyperparameter search was conducted during the training phase of each model through Optuna, with 10 trials, optimized on minimizing the evaluation loss during training using the search space: learning rate: $\{1e–5, 2e–5, 3e–5, 4e–5, 5e–5\}$, warm-up ratio: $\{0, 0.1\}$, batch size: $\{2, 4\}$, epochs: $\{2, 3, 4, 5\}$. The optimal hyperparameters were found based on grid search.

Following Tekiroğlu et al. (2022), we used the special tokens ‘<hatespeech>’ and ‘<counternarrative>’ , which were added directly before the HS and CN text, respectively. An instance of the training data is provided below:

(2) <hatespeech> Women must not have the right to choose. <counternarrative> Women should have equal rights in a free society, without discrimination. < |endoftext| >

Once the model was fine-tuned, the next step was to generate counter narratives for the test instances using the following decoding mechanisms: beam-search (Li et al., 2016), top-p (Holtzman et al., 2020), top-k (Fan et al., 2018), and a combination of the latter two, top-pk. The longest HS-CN pair in the dataset consists of 156 tokens, which corresponds to the token limit for the generation of CNs. The generated counter narratives were then compared to the gold CNs from the dataset using ROUGE-L (Lin, 2004) and BLEU-1 (Papineni et al., 2002) scores as automatic evaluation metrics. The ROUGE-L metric measures the similarity of text generated by a model compared to reference text based on the longest common subsequence between them, whereas BLEU-1 measures the overlap of unigrams (individual words) between the generated text and reference text. Since beam-search, which is designed to pick the most-likely sequence, yielded more generated texts compared to, e.g., top-k, which would often not generate a response to the HS instances in the test set, we proceed with beam-search as the decoding mechanism.

4.2 Strategies for incorporating author profiling information

For the strategies leveraging author profiling information, we continue with a subset of the data that we considered relevant for CN personalization. This subset was also split into 80-10-10 for training, validation, and testing, respectively.

Incorporating profiling information into HS instances in the test data (GPT-2-test) The test set was enriched with age and gender profiling aspects. An example of a test instance is provided below:

(3) <hatespeech> “Gay men are all perverts.” written by a woman older than 25 years. < |endoftext| >

This first experiment incorporated profiling information solely in the test set to see how well GPT-2 can adapt to extra information provided after fine-tuning.

Incorporating profiling information into HS instances in the training data (GPT-2-train) This experiment follows the idea by Chung et al. (2021). The authors create a dataset by combining existing counter narratives with factual information, resulting in the creation of more knowledge-grounded counter narratives. Hence, in this experiment, we include profiling characteristics of hate speech authors already in the fine-tuning process to explore whether the model can pick up on the profiling aspects and generate counter narratives that would be more targeted towards the author. The profiling information was included following the same technique, i.e., after the special token '<personalinformation>' as additional information to the HS-CN pair. An example of the extended dataset used for fine-tuning can be found below:

(4) <hatespeech> Immigrants steal our jobs. <personalinformation> written by a woman younger than 25 years. <counternarrative> Isn't it better to have people working hard and doing a good job, regardless of where they come from? < |endoftext| >

We examined different ways of phrasing profiling aspects. One example was including the information at the beginning of a sentence, i.e., "A woman younger than 25 years says: ...". However, this approach caused confusion in the generated sentences such that they, e.g., directly involved the same formulation at the beginning of the generated CNs making them very generic and repetitive. The phrasing showed in Example (4) yielded better CN generations, so we continued with the phrasing "written by a [gender] older/younger than [age] years" for the rest of the experiments.

Incorporating detailed profiling information into HS instances in the training data (GPT-2-detailed-train) Chung et al. (2021) incorporated extended factual information in their experiments using external knowledge sources. In line with their approach, we conducted experiments by adding more detailed profiling information to the training data. We used an automatic method of doing so to facilitate and increase the diversity of profiling aspects. Hence, we made use of OpenAI's[6] API, which employed the GPT-3.5 davinci model accessible during the conduction period of this research, and instructed it to do the following, given HS author's age and gender information previously obtained through the usage of TEXTGAIN (with one-shot prompting):

"Create a profile for a person of the following age and gender who wrote the following hate speech comment. The profile should include something about the time period they grew up in and the culture they are part of. Write maximally 2 sentences."

We set GPT-3.5's temperature parameter to 1 to create more creative and varied responses. Given the proficiency of GPT-3.5 in prompt prefixing (Brown et al., 2020), an exemplar of a more comprehensive profile was incorporated, taking into account factors such as time and culture, to provide the model with valuable learning material. We considered that the time period someone grew up in and the culture they are a part of would indicate underlying mechanisms of their attitude toward the individual and groups they are targeting in their hate speech. The example is as follows:

(5) Age: 55
Gender: male
HS: Women do nothing but spend their husbands money.
Profile: Michael grew up in the 1980s, a time when traditional gender roles were still deeply ingrained in society. He hails from a rural community where conservative values and expectations were prevalent.

These profiles were generated for each hate speech instance in the training data and added in the same way as described in the previous strategy, i.e., after the special token '<personalinformation>'. Below is an example from the training data:

(6) <hatespeech> A woman's place is with her husband or his family. <personalinformation> written by Paul who grew up in a society heavily influenced by traditional gender roles, where he was taught to believe that men were the

---

[6] https://openai.com/

decision makers while women should be subservient and obey the will of their husbands.<counternarrative> There are many cases where the wife works and provides for the family, making a valuable contribution. It is a source of pride for modern men to be married to these women.< |endoftext| >

Incorporating profiling information into the generation of personalized CNs using GPT-3.5 (GPT-3.5-personalized) In this approach, we experimented with GPT-3.5 for generating personalized CNs to explore its capability and effectiveness in tailoring responses specifically to individual hate speech authors.

We utilized GPT-3.5 with few-shot prompting to generate personalized counter narratives, taking into account the existing gold CNs. The temperature parameter here was also set to 1 to create more diverse CNs. The exact instruction was as follows:

"Your task is to take the following counter narrative and make it more personalized for a person of demographics described below."

The examples for the few-shot prompting looked like the following:

(7) Gold CN: I hate to break it to you, but women are not cunts. Feminists and humanists are fighting for the same rights that they hope to have in a thousand years. Hating women is the problem, not the solution.

Target of CN: Older man

Personalized CN: I understand that you may have grown up in a different time when attitudes towards women were different, but it's important to recognize that language like "cunt" is not acceptable and only serves to perpetuate harmful stereotypes and attitudes towards women. Feminists and humanists are not trying to take away your rights or make you feel inferior; rather, they are fighting for equality and the recognition that women should have the same rights and opportunities as men. This fight for equality is not new, and it is not going away anytime soon.

(8) Gold CN: The media are full of the rumours of rape and abuse committed by Muslim. The problem is not about Islam. It is about violence.

Target of CN: Younger woman

Personalized CN: I understand that the media can be overwhelming and that you may feel frustrated or angry about the topic of rape and religion. It can be difficult to sort through the rumors and misinformation. It's important to approach these issues with a critical eye and an open mind, and to recognize that not all Muslims are violent or abusive. The problem of violence is not unique to any one religion or culture. It affects people of all backgrounds and faiths, and we should work together to address it in a constructive and compassionate way.

An example of the generated personalized CN given the gold CN and demographics can be found below:

(9) Gold CN: Feminism is about the empowerment of women and having equal rights, what's not to like?

Target of CN: Younger man

Personalized CN: As a man, you have the power to contribute to an equal and equitable society for everyone—regardless of gender. Feminism is about more than just women gaining equal rights—it's about creating a world where everyone can be respected and safe. You have an opportunity to help create and enable this change and make sure that all voices can be heard and emphasized. It's about creating an environment where gender equality is a reality, not just a theoretical concept. Your support and participation can and will make a difference in the lives of those around you.

Incorporating personalized CNs in the training data (GPT-2-personalized-train) This strategy incorporates personalized CNs generated using GPT-3.5 as described in the previous paragraph for fine-tuning the GPT-2 model. Exclusively training the model with the gold CNs, even with the incorporation of profiling aspects, could result in the model compressing the inputs without effectively discerning the presence or absence of profiling information. By ensuring that the training data contains personalized counter narratives tailored to the given profiles, the GPT-2 model can potentially acquire the necessary knowledge to generate more targeted and engaging counter narratives. Accordingly, the training data looked like the following:

(10) <hatespeech> Women must not have the right to choose.<personalinformation> written by a man older than 25 years. <counternarrative> It can be difficult to adjust to a changing society, but it's important to recognize that gender equality and freedom are not just ideals, they are human rights that should be respected and upheld at all times. Women should have the same rights and opportunities as men, without any discrimination or prejudice. Older generations can help foster a culture of equality and respect by acknowledging that everyone deserves to be seen and treated equally, regardless of gender or any other background.< |endoftext| >

5 Evaluation

We report both manual and automatic evaluation of generated CN responses. We realize that automatic evaluation metrics may not reflect the personalization aspects of the CNs since we compare them to the gold CNs from the dataset. Accordingly, we also perform a manual evaluation, which ranks generated CNs with respect to each other and with respect to the gold CNs based on how personalized and engaging they are.

5.1 Automatic evaluation

We replicated the study by Tekiroğlu et al. (2022) on the 80-10-10 partitions of the complete dataset achieving a score of 0.0752 for ROUGE-L and 0.0836 for BLEU-1. The replication results demonstrate a notable discrepancy in the overlap metrics compared to the original study, which could be due to the fact that we did not exactly replicate the procedure described in (Tekiroğlu et al., 2022). In their paper, they generate five CNs for each HS and only select the CN with the highest ROUGE-L and BLEU score among the five to be evaluated. Tekiroğlu et al. (2022) reported an overlap score of 0.2195 for ROUGE-L and 0.2132 for BLEU-1 with beam-search as the decoding mechanism. We realize that direct comparison is not accurate since the training, validation, and test set distributions are not exactly the same, however, we wanted to see whether the results obtained through replication would be in line with the original study.

After achieving comparable results by fine-tuning GPT-2, we proceed with a subset of the data that we deemed relevant for CN personalization. The results of the different strategies exploring how to best generate personalized CNs are presented in Table 2.

Strategy	ROUGE-L	BLEU-1
GPT-2-test	0.0666	0.0776
GPT-2-train	0.0741	0.0842
GPT-2-detailed-train	0.0699	0.0728
GPT-3.5-personalized	0.1803	0.1618
GPT-2-personalized-train	0.0925	0.1411

Table 2: Results in terms of ROUGE-L and BLEU-1 score with beam-search as the decoding mechanism for the GPT-2 model.

The counter narratives generated using GPT-3.5 show the highest results in terms of the automatic evaluation metrics used. The results align with the results of the manual evaluation presented in Section 5.2.

5.2 Manual evaluation

The manual evaluation follows the idea by Haghighatkhah et al. (2022). We selected a subset of 30 HS instances and present every generated CN from the experiments as well as the gold CNs as a ranking task (see Appendix A for an example). We employed three master's students (two females and one male) with experience in hate speech-related tasks and asked them to rank the presented CNs from most (1) to least (7) personalized and engaging. There were seven CNs to be ranked for each hate speech instance—the gold CN, the CNs generated obtained through replicating the approach by Tekiroğlu et al. (2022), and the CNs generated by each of the five strategies described in Section 4.

To compute the pairwise inter-annotator agreement we use the Spearman correlation of their rankings. The average over all three pairwise agreement scores is 0.77.

To assess the relative performance of different CN generation methods, we adopt a methodology inspired by ranked-choice voting, specifically the Schulze voting system (Schulze, 2011). The notation $d[X, Y]$ is introduced, which represents the frequency with which a counter narrative generated by method X is preferred over a counter narrative generated by method Y across all hate speech instances and annotators. By calculating these pairwise preferences, a matrix is constructed that indicates the frequency of one strategy being preferred over another. The results are presented in Table 3.

Table 3 shows that the counter narratives generated with GPT-3.5 after instructing it to make them more personalized consistently attained higher ranking, even higher than the CNs created by experts, which secured the second highest ranking. The third highest ranking was achieved by the counter narratives produced during the replication of Tekiroğlu et al. (2022)'s study. This can be partially influenced by the fact that we used more training data than for the other strategies. The findings suggest that GPT-2 faces difficulties when integrating profiling information, such as age and gender, and particularly when tasked with incorporating more detailed information, it struggles to produce coherent CNs. One potential factor contributing to this challenge may be the utilization of beam-search as the decoding mechanism, which has a tendency for generating repetitive outputs. To improve the strategies for incorporating profiling characteristics, future research could investigate alternative decoding mechanisms and more fine-grained parameter optimization strategies.

6 Discussion

The utilization of GPT-2 was motivated by selecting one of the most promising outcomes from the aforementioned study (Tekiroğlu et al., 2022). The objective was to employ this model to explore various approaches for incorporating profiling information about hate speech users, aiming to identify an optimal method for personalizing counter narratives. The findings of the conducted experiments revealed that GPT-2 encounters challenges when attempting to generate personalized counter narratives. This pattern persisted throughout the various strategies employed for incorporating profiling aspects of HS authors. These findings suggest that further investigation and refinement of the strategies are necessary to enhance GPT-2's ability to generate more personalized responses to hate speech.

The results from both the automatic metrics and the manual evaluation reinforce the notion that employing a more advanced model exhibits a greater potential for transforming generic counter narratives into personalized and engaging ones. Leveraging the strong prompt-prefixing capability of GPT-3.5, we successfully generated personalized and engaging counter narratives indicating that, even with simple prompts, GPT-3.5 can effectively generate targeted and effective counter narratives. In the future, more detailed prompt engineering can potentially improve the personalization of generated CNs.

Factors such as removing items where age and gender are not relevant could have a significant impact on the results of this research since it also reduced the number of HS instances for training and testing. However, our main focus was finding the optimal approach to incorporate profiling information for personalization of CNs. Therefore, we conducted experiments using examples where age and gender potentially played a significant role (e.g., a man criticizing women) to observe the impact of each strategy, which provided valuable insights. Once the best strategy is identified, the next step would involve applying it to more hate speech examples where age and gender could be relevant, paving the way for real-world applications.

Given our focus on GPT, it would be intriguing to conduct a comparative analysis to assess the performance of alternative models such as BART (Lewis et al., 2019) or T5 (Raffel et al., 2020) and recent LLMs such as LLaMA (Touvron et al., 2023) and Koala (Geng et al., 2023) when injected with profiling information.

In line with Zeng et al. (2019)'s evaluation approach, the usage of personalized counter narratives generated with GPT-3.5 to combat online hate speech can additionally be evaluated by measuring user engagement, such as the number of likes or replies, to determine if there is a notable preference for personalized CNs.

7 Conclusion

The present study demonstrated that counter narratives can be made more personalized and engaging using the profiling characteristics of hate

	Gold-CN	Replication	GPT-2-test	GPT-2-train	GPT-2-detailed-train	GPT-3.5-personalized	GPT-2-personalized-train
Gold-CN	–	72.2	91.1	95.6	95.6	14.4	93.3
Replication (Tekiroğlu et al., 2022)	27.8	–	70.0	68.9	76.7	12.2	72.2
GPT-2-test	8.9	30.0	–	63.3	76.7	5.6	61.1
GPT-2-train	4.4	31.1	36.7	–	60.0	3.3	51.1
GPT-2-detailed-train	4.4	23.3	23.3	40.0	–	1.1	44.4
GPT-3.5-personalized	85.6	87.8	94.4	96.7	98.9	–	100.0
GPT-2-personalized-train	6.7	27.8	38.9	48.9	55.6	0	–

Table 3: The pairwise preferences from the manual evaluation in %. The values highlighted in green correspond to the percentage of one strategy (row) being preferred over another strategy (column). If this is not the case, the values are highlighted in red.

speech authors. Whilst this objective is possible with GPT-3.5 through simple prompting with profiling information, the personalization seems to be harder for GPT-2 throughout the various strategies explored for incorporating profiling aspects. Through incorporating author profiling information, we attempted to bridge the gap of generating more personalized counter narratives. Small glimpses into relevant findings regarding the incorporation of user information are a promising basis for researchers interested in focusing on the generation of less generic counter narratives, or in general, combating hate speech effectively. Ultimately, these findings shed light on the complex interaction between societal challenges such as hate speech in the digital realm, and the computational methods employed to tackle them. By examining the strengths and limitations of computational methods in addressing such challenges, we can strive towards more comprehensive and impactful strategies to promote a safer and more inclusive online environment.

8 Limitations

This study has potential limitations. The author profiling information relies on the TEXTGAIN tool, which may not be completely accurate in predicting demographic attributes. We also acknowledge that gender is not a binary matter, but TEXTGAIN does not provide more fine-grained gender classes. In future work, we will explore other strategies for identifying profiling characteristics of HS authors. Additionally, during manual evaluation, we only considered personalization and engagement, while other factors such as grammaticality and semantic correctness can also impact the effectiveness of generated CNs.

Acknowledgments

We would like to thank Helena Bonaldi for providing insights into the code used in their study. We would also like to thank anonymous reviewers for their feedback that helped improve this paper. All remaining errors are our own.

References

Susan Benesch. 2014. Countering Dangerous Speech: New Ideas for Genocide Prevention. *Working Paper*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marco Casavantes, Mario Ezra Aragón, Luis C. González, and Manuel Montes-y-Gómez. 2023. Leveraging Posts' and Authors' Metadata to Spot Several Forms of Abusive Comments in Twitter. *Journal of Intelligent Information Systems*, pages 1–21.

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. *CoRR*, abs/1910.03270.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. ACL.

Javid Ebrahimi and Dejing Dou. 2016. Personalized Semantic Word Vectors. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1925–1928, New York, NY, USA. ACM.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. ACL.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. 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*, pages 3226–3240, Online. ACL.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering Online Hate Speech*. Unesco Publishing.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A Dialogue Model for Academic Research. Blog post.

Pantea Haghighatkhah, Antske Fokkens, Pia Sommerauer, Bettina Speckmann, and Kevin Verbeek. 2022. Story Trees: Representing Documents using Topological Persistence. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2413–2429, Marseille, France. ELRA.

Lisa Hilte, Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2023. Who are the haters? A corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual Syntactic Variation Over Age and Gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. ACL.

Pan Li and Alexander Tuzhilin. 2019. Towards Controllable and Personalized Review Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3237–3245, Hong Kong, China. ACL.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81.

Ilia Markov and Walter Daelemans. 2021. Improving Cross-Domain Hate Speech Detection by Reducing the False Positive Rate. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 17–22, Online. ACL.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring Stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech Detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. ACL.

Johanna Monti, Valerio Basile, Maria Pia Di Buono, Raffaele Manna, Antonio Pascucci, and Sara Tonelli. 2022. Proceedings of the Second International Workshop on Resources and Techniques for User Information in Abusive Language Analysis. Marseille, France. ELRA.

Gwenn Schurgin O'Keeffe, Kathleen Clarke-Pearson, et al. 2011. The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics*, 127(4):800–804.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Eli Pariser. 2011. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK.

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, pages 4755–4764, Hong Kong, China. ACL.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Francisco Rangel, Gretel Liz de la Peña-Sarracén, María Alberta Chulvi-Ferriols, Elisabetta Fersini, and Paolo Rosso. 2021. Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum*, pages 1772–1789. CEUR.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection Using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. ACL.

Markus Schulze. 2011. A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method. *Social choice and Welfare*, 36(2):267–303.

Serra Sinem Tekiroğlu, 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 the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. ACL.

Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020a. 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, Yi-Ling Chung, and Marco Guerini. 2020b. 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. ACL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LlaMA: Open and Efficient Foundation Language Models.

Wenhuan Zeng, Abulikemu Abuduweili, Lei Li, and Pengcheng Yang. 2019. Automatic Generation of Personalized Comment Based on User Profile. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 229–235, Florence, Italy. ACL.

A CN examples generated using each of the strategies for a given HS instance

HS: I wish we lived in a world in which only men could be doctors, lawyers, firemen, and teachers. (written by a man older than 25 years)
Gold-CN: In the U.S., a doctor, a lawyer, a teacher are some of the most highly valued professions. Women working as doctors and lawyers and teachers are highly respected professionals as well, right?
Replication: We need more women involved at all levels of society to build what is possible for the future generations! For example: not limiting participation by gender equality criteria like 'gender'. Society can live with both genders equally contributing fully together toward creating something great? And I am sure you think that's good enough??? Just stop confusing people instead.
GPT-2-test: Do you agree with the statement? Do any of these statements have that sexual orientation is just an issue to keep people from becoming professionals or entrepreneurs because there aren't enough employees around for them: LGBT+ communities are everywhere now... So where do they come out next.. It's hard but worth it! Who knows how long this will lead us back into trouble??? How can anyone make such false claims about what happens inside social media!? #LGBTFeministLIVESJUMP — Jen Purdy (@JenProuder)
GPT-2-train: Just about any decent human being would have said something like this: "Why do millions of people want to live their lives just as they did before?
GPT-2-detailed-train: The word "crisis" is often used as opposed with respect merely for problems ranging from violence against women towards all racial groups or ethnicities etc.: But that doesn't mean you don right? I'm saying how do things here-it's different than what everyone else thinks! You can say anything YOU want around this forum without being afraid it wont any more:You shouldn't.
GPT-2-personalized-train: You must make sure there are no problems with someone else's work - at least not if you have children of either sex or any other gender-related issues here on the UK Blogs! For practical matters please don't just say how much time your child would like to spend researching this subject before deciding for yourself whether that is appropriate: such research may become controversial; as well should debate about why some people prefer sexual health care without regard their differences too often – when they benefit more from it rather do others? I believe all lives depend upon each one being valued equally
GPT-3.5-personalized: I understand that you may have grown up in a different era when some professions were seen as being more suitable for women than men. It's important to recognize, however, that this is no longer the case and that women should be seen as equal players in the fields of medicine, law, and education. Highly skilled women working in these professions should be respected and appreciated for their hard work and dedication. Women can and should be just as successful and celebrated as their male counterparts. No one should be held back from opportunities or hindered by gender roles.