

Research Master Thesis

Automatic Generation of Personalized Counter Narratives Based on User Profile

Mekselina Doğanç

*a thesis submitted in partial fulfilment of the
requirements for the degree of*

MA Linguistics
(Human Language Technology)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities



Supervised by: Dr. Ilija Markov
2nd reader: Prof. Dr. Piek Th. J. M. Vossen

Submitted: June 30, 2023

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. Manual creation of CNs allows for personalization but previous studies have shown that computational models often generate CNs that are dull and generic and therefore do not resonate with hate speech authors. This thesis explores the potential of utilizing advanced language models, specifically GPT-2 and GPT-3.5, to enhance the automatic generation of personalization of counter narratives against hate speech. The objective is to investigate whether integrating profiling information about HS authors and employing more powerful language models can transform generic and dull counter narratives into engaging and personalized responses.

The research methodology involves a series of experiments and evaluations using a qualitative HS-CN dataset. The initial replication of a study by [Tekiroğlu et al. \(2022\)](#) establishes a baseline, while subsequent experiments focus on incorporating author profiling aspects, such as age, gender, and more detailed profiling information into fine-tuning and testing. Automatic metrics and manual evaluations are employed to assess the personalization and engagement levels of the generated counter narratives.

The results indicate that GPT-3.5, with its advanced capabilities and larger parameter size, outperforms GPT-2 in terms of generating personalized and engaging counter narratives. The inclusion of author profiling aspects demonstrates a significant impact on enhancing personalization. The challenges, opportunities, and future directions for incorporating user information into CN interventions are discussed as well. The code is publicly available to help advance the research on counter narrative generation.

Declaration of Authorship

I, Mekselina Doğanç, declare that this thesis, titled *Automatic Generation of Personalized Counter Narratives Based on User Profile* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 30.06.2023

Signed: 

Acknowledgments

I would like to state my infinite gratitude to my supervisor, Dr. Ilia Markov, who did not only motivate me to go out of my comfort zone and work with natural language generation but was also always present when I needed his help and guidance. This thesis would not have been possible without his trust and assistance. It was essential for me to work on such a socially impactful topic and I hope to further do so in the future. I would like to address special thanks to Helena Bonaldi for sharing the code of her study, hence supporting me tremendously and showing that NLP and NLG can thrive so much more under the roof of open science. Furthermore, I would like to express my deepest appreciation to my friends Lukas Schießer and Polina Tsvilodub for their invaluable support. You two play a huge role in my academic path, once a Coxi, always a Coxi! Finally, I would like to thank my friends Elena, Vicky, and Giorgio for their immediate support with the annotation. In particular, I want to thank Elena, my roomie, who kept up with my highs and lows throughout the past 5 months. I feel very blessed to be around people who unconditionally support and inspire me.

List of Figures

2.1	Guidelines from https://getthetrollsout.org/stoppinghate on how to counter hate speech online.	8
2.2	A visualization of the transformers architecture taken from the original paper “Attention is All You Need” (Vaswani et al., 2017)	15
3.1	A visualization of the data collection pipeline taken from Fanton et al. (2021): “The author-reviewer in the loop configuration. The author module produces HS/CN candidates and the reviewer(s) validates and eventually post-edits them. At each loop new examples are added to training data and the author is fine-tuned from scratch.”	20
3.2	The distribution of HS-CN pairs per target in the MultiCONAN dataset. .	21
3.3	The distribution of HS-CN pairs per author profile in the modified dataset. .	27
3.4	Excerpt from the annotations where the column “COUNTER_SPEECH” has 7 rows with different (generated) CNs to rank from most (1) to least (7) personalized and engaging.	33

List of Tables

3.1	The number of HS-CN pairs per target of HS. The targets that don't fit in any of the target groups are labeled as "other".	21
3.2	The number of HS-CN pairs per author profile in the modified dataset. .	27
4.1	Results in terms of ROUGE-L and BLEU-1 score with beam-search as the decoding mechanism.	35
4.2	The pairwise preferences from the manual evaluation in %	36

Contents

Abstract	i
Declaration of Authorship	iii
Acknowledgments	v
List of Figures	vii
List of Tables	ix
1 Introduction	1
2 Background & Related Work	5
2.1 What is Hate Speech?	6
2.2 What are Counter Narratives?	7
2.2.1 Counter Narrative Generation	9
2.3 Author Profiling	11
2.4 Model	13
2.4.1 Transformers	13
2.4.2 Transformers and GPT	16
3 Methodology	19
3.1 Data	19
3.2 Replication	21
3.3 Strategies for Incorporating Profiling Information	27
3.3.1 Incorporating profiling information into HS instances in the test data (GPT-2-test)	27
3.3.2 Incorporating profiling information into HS instances in the training data(GPT-2-train)	28
3.3.3 Incorporating detailed profiling information into HS instances in the training data (GPT-2-detailed-train)	28
3.3.4 Generating personalized CNs and integrating them into the HS instances in the training data (GPT-2-personalized-train) & (GPT-3.5-personalized)	29
3.4 Evaluations	31
3.4.1 Automatic Evaluation	31
3.4.2 Manual Evaluations	33

4 Results & Analysis	35
4.1 Automatic Evaluation	35
4.2 Manual Evaluation	36
5 Discussion	41
6 Conclusion	45
A Appendix A	47
A.1 Annotation guidelines for manual evaluation	47

Chapter 1

Introduction

“No one is born hating another person because of the color of his skin, or his background, or his religion. People must learn to hate, and if they can *learn* to hate, they can be *taught* to love.” [Mandela \(2008\)](#)

Hate speech (HS) is spreading through various channels, such as social media, online forums, and extremist websites [\(Chaudhary et al., 2021\)](#). The disinhibition effect, prevalent in online environments, diminishes social inhibitions and prompts individuals to engage in unrestrained and offensive behavior [\(Suler, 2004\)](#), thus facilitating the spread of hate speech. Exploiting the anonymity and wide reach of the internet, individuals can readily target vulnerable communities with their hateful messages and ideologies [\(Banks, 2010; O’Keeffe et al., 2011\)](#). Hate speech does not only affect minorities who usually don’t have the power to make their voices heard, and their discreteness leads to xenophobia (fear of the unknown)—someone who does not know enough about Roma people is more likely to portray them as beggars and criminals ¹. Hate speech also affects demographics and members of religious groups that have greater representations in society such as women and Muslims. It is the permanence and continuity of opinions and prejudices towards the representatives of different “groups” that also find responses in echo chambers online and propagate hate in a “positive feedback-loop” [\(Pariser, 2011\)](#). The positive feedback loop limits the exposure to alternative views, which disturbs the balance of different points of view [\(Pariser, 2011\)](#).

The spread of hate speech can be motivated by a variety of factors, including fear, ignorance, and prejudice [\(Banks, 2010\)](#). Some individuals may spread hate in an attempt to increase their own power or status, while others may do so out of a desire to belong to a particular group or community [\(Banks, 2010\)](#). Regardless of the reason, the spread of hate speech has serious consequences, leading to increased discrimination, violence, and division within society [\(Bilewicz and Soral, 2020; O’Keeffe et al., 2011\)](#). Hence, the topic of hate speech is not an issue limited to only one field but rather an interdisciplinary and international one. There are articles on hate speech from fields such as law, sociology, and communication [\(Banks, 2010; Gagliardone et al., 2015\)](#). While it is important to understand and address the root causes of hate speech in order to prevent its spread, it is also crucial to promote understanding and inclusion. As Nelson Mandela says in the above quote, love and hate are not inherent but can be taught.

One attempt to do so is through responses to hate speech that promote understanding and different perspectives, known as counter narratives. Identifying and deleting hate speech posts and comments or blocking the authors that write hate speech may cause

¹<https://www.dw.com/en/germany-discrimination-against-roma-and-sinti-on-the-rise/a-65173343>

more agitation and fury in those authors (Munger, 2017). If they really want to share their opinion, they will find a way and source to do so. Instead, tackling the situation with counter narratives can de-escalate the rage in authors and motivate them to re-evaluate what has been posted (Benesch, 2014; Gagliardone et al., 2015). NGOs work for exactly this purpose by hiring people to write counter narratives² however, there is a great imbalance between the number of hate speech shared every day and the responses written by NGO workers. In trying to fight this imbalance, responses can get repetitive and generic while also appearing dull and not engaging (Qian et al., 2019). On top of that, regarding the sensitivity of this topic, especially if an NGO worker feels a belonging to one of the target groups of the hate speech, responding can be very disturbing over time and affect the workers mentally.

Therefore, some natural language generation (NLG) studies have suggested automatic ways of counter narrative generation (Qian et al., 2019; Chung et al., 2019; Tekiroğlu et al., 2022). Natural language processing (NLP) has already been a great help for successfully identifying hate speech and classifying it as something to get rid of. However, the goal is not only to analyze and identify hate speech, NLP/NLG scientists have to go beyond and also create solutions for generating counter narrative. Tekiroğlu et al. (2022) from the lab Fondazione Bruno Kessler (FBK)³ whose study is also taken as a baseline for this thesis, has addressed this issue and demonstrated that the process of countering hate speech can also be enforced and supported through NLG. Even though these approaches are very promising for the purpose of combating the tremendous amount of hate speech rising every hour, they tend to overlook the fact that most of the counter narratives stay generic and not engaging. There has been a growing interest in the NLP and sociolinguistic communities regarding the profiling of individuals who create hateful content (Hilte et al., 2023), and the information pertaining to the identity of these hate speech propagators offers valuable insights into the motivations behind their dissemination of such content. This knowledge not only enhances our understanding of the phenomenon itself but also provides opportunities for enhancing hate speech detection systems through the incorporation of profiling and stylometric information (Markov et al., 2021). As an example, let's examine the following instance of hate speech along with two potential counter narratives aimed at addressing it:

(1) HS: Women must not have the right to choose (written by an older man).

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.

By incorporating the background information of the hate speech being authored by an older man, CN2 effectively addresses the hate speech by incorporating personalized elements (as highlighted in bold). In contrast, CN1 fails to achieve this level of

²<https://getthetrollsout.org/>

³<https://www.fbk.eu/en/>

personalization, resulting in a more generic and less engaging response.

However, it is important to note that the exploration of profiling hate speech propagators within the context of personalization of counter-narrative interventions remains an unexplored area of research. Consequently, the center of attention in my thesis is exploring different ways to make counter-narrative responses less generic, more personalized and engaging tailored to each hate speech author. This can potentially stop them from confining into isolated information communities and restricting them from seeing the vast array of other possibilities.

To bridge this gap, I conduct multiple experiments in this thesis encompassing several approaches. First, I replicate the work of (Tekiroğlu et al., 2022) by utilizing GPT-2 (Radford et al., 2019) for counter narrative generation. Additionally, different strategies are explored, such as injecting profiling information exclusively in the test set, fine-tuning GPT-2 with profiling information, employing more detailed profiling, and incorporating personalized counter narratives generated using GPT-3.5 (Ouyang et al., 2022). These experiments aim to investigate the effectiveness of profiling information in achieving personalized and engaging counter narratives. The findings demonstrate that leveraging advanced language models like GPT-3.5 enables the creation of highly focused counter narratives targeted to the specific hate speech authors.

The thesis is structured as follows: Chapter 2 introduces relevant studies that have dealt with hate speech and counter narratives, as well as studies about author profiling. It will also explain the mechanism behind GPT-2 and GPT-3.5. In Chapter 3, the various strategies employed for incorporating profiling information into the generation of personalized CNs is explained and the results of these strategies are presented and analyzed in Chapter 4. Chapter 5 discusses these results whereas the last chapter addresses the impact of the findings.

Chapter 2

Background & Related Work

Combating hate speech (HS) is a long-due topic that requires a multi-faceted approach, involving the efforts of governments, businesses, community organizations, and individuals (Banks, 2010; Gagliardone et al., 2015). Raising public awareness and educating people about the impact is one approach along with communities coming together to speak out against hate speech and promote diversity, tolerance, and understanding. Certain countries also undertake legal actions such as paying fines or even imprisonment¹. Another way are social media policies—policies to prohibit hate speech and remove it when it is reported (Chetty and Alathur, 2018). However, since there is a fine line between free speech and hate speech, monitoring the deletion or blocking is challenging and may lead to over-blocking (i.e., excessively restricting or censoring online content or user accounts, going beyond what is reasonably necessary to protect the platform or users) (Markov and Daelemans, 2021).

According to the Pew Research Center (Duggan, 2017), 45% of Americans prioritize allowing free expression online, while a slightly larger share (53%) prioritize creating a welcoming and safe online environment. The special rapporteurs from the Office of the United Nations High Commissioner for Human Rights (OHCHR)² have endorsed the approach of “more speech” as a strategic response to hate speech (Bielefeldt et al., 2011). Hence, it is morally important to not enforce silence by blocking, instead “more speech” can be used, perhaps through creating encouraging, positive, and inclusive counter narratives (CNs). Indeed, hate speech can be effectively combated by employing counter narratives (Qian et al., 2019; Tekiroğlu et al., 2020). It is important to understand what forms hate speech and counter narratives can take to further acknowledge how effective this method can be. However, this thesis looks beyond the identification of hate speech and counter narratives but specifically deals with the generation of the latter. While it is important to have an automatic way of doing so, it is even more important to do so in a qualitative manner to not defeat the whole purpose of addressing and, ideally, educating the hate speech authors. The authors of hateful content can only be properly addressed if the computational model used for generating counter narrative has access to profiling information about the users and can integrate this information into the generation. Whether this profiling information can make the generation more personalized or not, is explored in this thesis with different experiments.

The experiments in this study utilize the language models GPT-2 (Radford et al.,

¹<https://www.legislation.gov.au/Details/C2016C00089>

²https://www.ohchr.org/en/ohchr_homepage

(2019) and GPT-3.5³. The architecture of the language models will be described in detail in this chapter. Furthermore, the subsequent sections aim to provide a comprehensive understanding of hate speech and counter narratives, highlighting the significance of automatic counter narrative generation in combating online hate speech. Additionally, the impact of incorporating profiling information in prior research and the computational approach employed for generating counter narratives in this thesis will be elucidated.

2.1 What is Hate Speech?

While many have a general idea about what hate speech entails, studies have shown that formulating a proper definition can be difficult. Which elements does hate speech contain? How subjective is hate speech? How do we include the speaker’s intent in a definition? All in all, many questions play a part in defining this term. Nevertheless, many academics have attempted to formulate a definition.

A review of different definitions given to the concept of hate speech concluded that there are different ways from which to approach a definition. These different approaches are: harm-based; which concerns the actual harms its receivers are subjected to, content-based; which focuses on the content of the slur, an approach based on intrinsic properties; focuses mainly on inherently derogatory, discriminatory, or vilifying types of speech, and lastly dignity-based; focuses mainly on the harm done to the dignity of the subject of the utterance (Anderson and Barnes, 2022). An example for each approach is given below.

For the harm-based approach, Brison (1998) defines hate speech as “speech that vilifies individuals or groups based on such characteristics as race, sex, ethnicity, religion, and sexual orientation, which (1) constitutes face-to-face vilification, (2) creates a hostile or intimidating environment, or (3) is a kind of group libel”. Parekh (2012) gives a content-based definition, saying that hate speech “expresses, encourages, stirs up, or incites hatred against a group of individuals distinguished by a particular feature or set of features such as race, ethnicity, gender, religion, nationality, and sexual orientation”. A definition based on intrinsic properties by Anderson and Barnes (2022) concludes that “hate speech is speech that is inherently derogatory, discriminatory, or vilifying”.

Lastly, Waldron (2012) formulates a dignity-based definition as follows: “[hate speech is] speech that undermines its target’s “basic social standing, the basis of [their] recognition as social equals and as bearers of human rights and constitutional entitlements”. A definition that combines all these aspects would be a better attempt at acting as the most inclusive definition.

Finally, the European Union struggled with the difficulty of describing hate speech in criminal law. Several definitions exist across multiple articles and policies. Article 20 of the International Covenant on Civil and Political Rights of 1966 solely includes national, racial, or religious hatred⁴. This is similar to Article 4 of the International Convention on the Elimination of Racial Discrimination from 1965, which has “any group of persons of another color or ethnic origin” as subjects of hate speech⁵. A broader definition of the subject was constructed for the definition guiding the current study. This is more in

³<https://openai.com/>

⁴<https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>

⁵<https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial>

line with the Committee of Ministers of the Council of Europe Recommendation (No R 97(20)) on hate speech, which describes hate speech as “any form of hatred based on intolerance”, which includes hostility against minorities, migrants and people of immigrant origin⁶.

In order to keep one guiding definition, the Council of Europe later decided on one, clear definition in their combat against certain forms of expressions. In a Framework Decision, hate speech is defined as “publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, color, religion, descent or national or ethnic origin”⁷.

Taking this definition into consideration, let’s have a look at how counter narratives are explained.

2.2 What are Counter Narratives?

Counter narratives offer feedback to any hate speech user in a friendly manner using factual arguments and can stop any starting trend of a negative spiral of hate (Tekiroğlu et al., 2020).

Defining counter narrative has proven difficult, as it can take many forms. Several researchers have proposed definitions of counter narratives, and have written about strategies of counter narratives, the forms it can take, and their respective effectiveness.

One definition comes from Benesch et al. (2016): “counter narrative is defined in this study as a response that takes issue with hateful, harmful or extremist content”. The definition states that counter narrative is a direct response to hate speech and seems to include more about the “reason” behind counter narratives, namely that the user takes issue with the hate speech. But, how can we be sure of the quality of the counter narrative so that we don’t push counterproductive counter narratives? Consider the following example given by Chung et al. (2019):

(1) HS: “I hate Muslims. They should not exist.”

CN1: “Muslims are human too. People can choose their own religion.”

CN2: “You are truly one stupid backwards thinking idiot to believe negativity about Islam.”

CN1 would be an appropriate and non-negative response to the given HS whereas CN2 would escalate the situation even more. There have to be some standards set for a good counter narrative because only then can it have an educating and de-escalating effect.

In the same article by Benesch et al. (2016), they found eight main strategies to counter narratives: “1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions, 3) warning of possible offline and online consequences of speech, 4) identification with original speaker or target group, 5) denouncing speech as hateful or dangerous, 6) use of visual media, 7) use of humor, and 8) use of a particular tone, e.g. an empathetic one.” (Benesch et al., 2016). As they point out, it is possible (if not common) for a user to use multiple of these strategies. Another

⁶<https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021DC0777>

⁷<https://eur-lex.europa.eu/legal-content/EN/LSU/?uri=celex:32008F0913>

definition by Benesch et al. (2016) focuses more on the effect of counter narratives: “We define successful counter narratives in two ways. The first is speech (text or visual media) that has a favorable impact on the original (hateful) user, shifting his or her discourse if not also his or her beliefs. [...] The second type of success is to positively affect the discourse norms of the “audience” of a counter narrative conversation.”

Counter narratives are already applied by initiatives such as “Get The Trolls Out”⁸ aimed at combating and raising awareness about hate speech, particularly in online environments. It focuses on countering discriminatory narratives and promoting inclusive dialogue. The initiative employs various strategies, including monitoring and reporting hate speech incidents, advocating for responsible online behavior, and providing resources and guidelines for individuals and organizations to effectively address hate speech. Figure 2.1 displays an excerpt from their website with a recommendation on how to effectively combat hate speech.

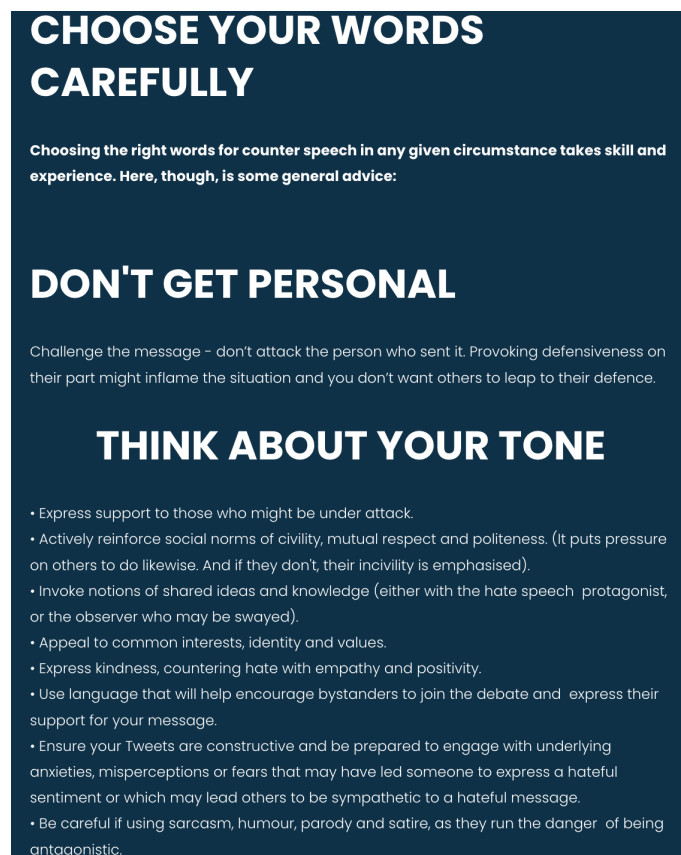


Figure 2.1: Guidelines from <https://getthetrollsout.org/stoppinghate> on how to counter hate speech online.

Generally said, counter narratives aim to address the underlying biases, misinformation, or prejudice embedded in hate speech, while fostering understanding, empathy, and tolerance among online users. The importance of counter narrative generation lies in its potential to shift the narrative, steer conversations toward productive discourse, and provide a voice to individuals and communities affected by hate speech. By actively challenging and debunking harmful ideas and narratives, counter narratives contribute

⁸<https://getthetrollsout.org/stoppinghate> on how to counter hate speech online.

to the creation of a more inclusive, respectful, and empathetic online environment.

2.2.1 Counter Narrative Generation

Counter narrative generation offers several key benefits in the fight against online hate speech. Firstly, it empowers individuals and communities to reclaim their online spaces by providing them with a means to express their perspectives, share personal experiences, and challenge discriminatory ideologies. This empowerment fosters resilience, self-advocacy, and community-building among those affected by hate speech (Benesch et al., 2016). Secondly, counter narrative generation plays a crucial role in educating and enlightening those who may be influenced by hate speech or exposed to biased narratives. By 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 (Chung et al., 2021a). Furthermore, counter narrative generation serves as a preventative measure, intercepting the potential escalation of hate speech into real-world harm. By promptly and effectively countering hate speech with reasoned and empathetic responses, it can disrupt the spread of harmful ideologies, discourage the normalization of discriminatory beliefs, and mitigate the potential for offline harm (Tekiroğlu et al., 2020).

In recent years, advances in artificial intelligence (AI) and natural language generation (NLG) have facilitated the development of computational approaches for automatic counter narrative generation. These approaches leverage linguistic and contextual cues to generate persuasive and educating responses to hate speech, thereby assisting individuals and communities in their efforts to combat online hatred (Chung et al., 2021b).

The availability of good datasets is the first big step toward the goal of generating qualitative counter narratives. For this purpose, the lab Fondazione Bruno Kessler (FBK)⁹ curated multiple high-quality and high-quantity datasets consisting of HS-CN pairs¹⁰ (Chung et al., 2019; Fanton et al., 2021; Chung et al., 2021a). The quality is assured through niche sourcing—a niche group of experts in data collection e.g., NGO operators who have the experience of generating counter narratives that will hinder the beginning of a negative spiral. For instance, MultiCONAN is one of the datasets consisting of HS-CN pairs for English comprising several hate targets collected with a human-in-the-loop approach (human input and expertise are incorporated into the process of gathering or labeling data) (Fanton et al., 2021). Examples of the dataset can be found below:

- (2) HS: Gays are just a bunch of disgusting animals. You should have the same disgust for them that you have for cancer people, small pox, and TB.
CN: This is just prejudice that is not based on facts.
- (3) HS: If women could have 100% control over the conception, they would have never had children.
CN: Women are free to have as many children as they want.

Example (2) targets members of the LGBT+ community whereas example (3) targets women.

⁹<https://www.fbk.eu/en/>

¹⁰The datasets can be found under the following link: <https://github.com/marcoguerini/CONAN>

The same lab also provides a dataset for a more knowledge-grounded counter narrative generation to fight the hallucination phenomena—a process involving a language model producing fictional or incorrect information that does not exist in the input data or the real world. Therefore, the authors reverse-engineer a dataset where they take the counter narrative that is already existing and pair them with facts (Chung et al., 2021a). An example of the dataset is as follows:

- (4) HS: Jews control the banks, media, hollywood, even the U.S. government.

Knowledge Retrieval: “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 as early as 1921-become one of Hitler’s favorite texts. Jews were accused of desecrating the sacramental bread used in communion, poisoning wells and spreading the plague.

CN: 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. As early as 1921-become one of Hitler’s favorite texts. 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.

The knowledge retrieval part captures relevant factual information and relationships from external knowledge sources, such as structured databases or knowledge bases which enables the model to utilize external knowledge during the generation of counter narratives effectively (Chung et al., 2021a).

The same authors introduce a new platform designed for NGO operators to monitor and analyze social media data, along with a tool for suggesting counter narratives (Chung et al., 2021b). The platform aims to improve the efficiency and effectiveness of operators’ efforts to address islamophobia by reducing the time required to generate counter narratives. The approach in developing the counter-narrative tool involved utilizing a retrieval-based suggestion engine that generated predefined suggestions for similar hate messages. Users were then instructed to rephrase the suggested counter narratives to enhance the diversity of the messages shared online. (Chung et al., 2021b) realize that this is a limitation and suggest considering the adoption of alternative methods that leverage advancements in neural text generation to generate more varied counter narratives.

Not only does the lab FBK provide qualitative datasets and a platform for countering hate speech, but also a study where they use different computational approaches to generate counter narratives automatically. One such study, and the main source for this thesis, is presented in Tekiroğlu et al. (2022). This study compares the performance of pre-trained language models (e.g., GPT-2, BERT, and T5) in generating counter narratives to fight hate speech online. The findings show promising results in generating diverse and coherent counter narratives so that e.g., criteria like “1) presentation of facts to correct misstatements or misperceptions, 2) pointing out hypocrisy or contradictions” by Benesch et al. (2016) are met. Nevertheless, further improvements are needed to address issues such as repetition and lack of contextual relevance. And to be able to have the desired effect of “favorable impact on the original (hateful) user, shifting his or

her discourse if not also his or her beliefs” (Benesch et al., 2016), the counter narratives need to be somewhat tailored to the hate speech author. However, to the best of our knowledge, profiling hate speech spreaders has not been explored in the context of personalization of counter-narrative interventions yet. To understand the effect that author profiling can have, the following section will explore previous studies where this effect was proven.

2.3 Author Profiling

Author profiling refers to the process of extracting relevant attributes and characteristics of individual authors from their linguistic expressions, social behavior, and demographic information. By considering factors such as age, gender, native language, personality traits, educational level, cultural background, ethnic background, and political orientation, author profiling offers a more comprehensive understanding of the context in which hate speech is generated. This personalized analysis allows us to delve beyond the surface level of offensive words and phrases and gain insights into the underlying motivations, biases, and intentions of the authors (Mishra et al., 2018). The personalized analysis enables more accurate identification and characterization of hate speech instances by accounting for the unique perspectives, backgrounds, and influences of individual authors (Mishra et al., 2018). By unraveling the demographic and psycholinguistic dimensions of authors, we gain a deeper understanding of the sociocultural factors that contribute to the propagation of hate speech, helping us to devise targeted interventions and policy measures.

Up to now, numerous studies have dedicated their attention to author profiling and its advantages, particularly in terms of comprehending the underlying factors behind different author’s online behavior and producing text specifically tailored to individual 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). Traditional approaches to hate speech detection have predominantly relied on lexical and syntactic features, focusing on the content of the messages themselves (Schmidt and Wiegand, 2017). 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 making counter narratives more personalized.

To take one study as an example, Johannsen et al. (2015) contributes to our understanding of how age and gender influence syntactic variation across languages. The study investigates how these demographic factors contribute to linguistic differences and whether such variations are consistent across languages or exhibit language-specific patterns. For this purpose, the researchers analyzed large-scale multilingual datasets comprising written texts from diverse sources and populations. The findings of the study reveal intriguing patterns of cross-lingual syntactic variation. Firstly, the researchers observe consistent age-related differences in syntactic structures across languages. Younger individuals tend to exhibit a higher usage of certain syntactic constructions compared to older individuals, indicating generational language change. This suggests that age plays a role in shaping language use and syntax, transcending linguistic boundaries. Secondly, the study uncovers gender-related differences in syntactic variation that are both language-specific and cross-linguistic. While some syntactic features show gender-based distinctions in all languages, others exhibit variations unique to specific languages.

These findings highlight the interplay between gender and language in shaping linguistic expression.

Another example is provided by [Ebrahimi and Dou \(2016\)](#). The authors propose a methodology to create personalized word vectors by leveraging the rich source of information available from social media platforms. The approach involves collecting user-generated text data, such as tweets or forum posts, and using this data to capture the specific semantic associations that each user has with different words. By considering the user's own linguistic context, the generated word vectors are personalized to reflect their unique semantic perspectives. The findings suggest that personalized word vectors have the potential to enhance various natural language processing applications by incorporating individual semantic variations.

In addition to the studies mentioned above, in the paper “Profiling Hate Speech Spreaders on Twitter Task at PAN 2021” ([Rangel et al., 2021](#)), one of the approaches utilized user profiling information, such as their profile description, number of followers, number of friends, and account creation date to profile individuals who spread hate speech on Twitter. This approach aimed to leverage these profiling information features to gain insights into the characteristics of hate speech spreaders and distinguish them from non-hate speech users. For example, the number of followers and friends of a user can reflect their social network and reach. Hate speech spreaders may have a higher proportion of followers or friends associated with hate speech-related accounts or communities. Or the textual information provided by users in their profile descriptions can offer valuable insights into their interests, beliefs, and potential biases. Analyzing the content of user descriptions can help identify hate speech-related keywords or themes. By considering user profiling information, this approach aimed to provide additional context and insights into the behavior and characteristics of hate speech spreaders on Twitter.

The study by [Casavantes et al. \(2023\)](#) deals with abusive language detection as well by leveraging posts' and authors' profiling information. The study aims to enhance the identification of abusive content by considering contextual information associated with individual posts and authors. Post-level profiling information includes features such as the length of the post, the presence of specific keywords, the number of hashtags or mentions, and the presence of offensive language. Author-level profiling information encompasses information about the author, such as their account age, follower count, and verified status. Using a machine learning approach, the authors train models to classify comments into different categories of abusive content, such as hate speech, offensive language, or cyberbullying. The models are trained using a labeled dataset where abusive comments have been manually annotated. The study evaluates the effectiveness of their approach by comparing the performance of models that utilize post and author profiling information against models that only consider textual content. The results demonstrate that incorporating profiling information significantly improves the detection accuracy of abusive comments across various categories. Furthermore, the authors analyze the importance of different profiling information features in identifying abusive content. They identify which aspects of post and author profiling information contribute most to the detection of specific abusive categories. This analysis provides insights into the relevance of different contextual factors in detecting abusive comments.

A conceptually similar work to my thesis 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 presents an approach that leverages user profile information, such as demographic data, interests, and preferences, to generate comments that align with the user's profile. Therefore, the authors collected and analyzed user profiles to extract relevant features, such as age, gender, location, and stated interests. Afterward, they trained a machine learning model 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 can then be used to generate personalized comments for new users based on their profile information. By inputting a user's profile data into the model, it generates comments that are tailored to the user's specific attributes, preferences, and interests. The generated comments are designed to be contextually appropriate and relevant, enhancing user engagement and fostering a more personalized user experience.

Yet another study by [Li and Tuzhilin \(2019\)](#) learns to capture user-specific patterns by leveraging user data and preferences for the aim of developing a review personalization model that takes into account user-specific information, such as historical preferences, demographic data, or previous reviews, to generate reviews that align with the user's personal tastes and preferences.

As evidenced by the various studies conducted, the incorporation of profiling and stylometric information in hate speech detection systems not only enhances their performance but also provides valuable insights into the phenomenon of hate speech. By gaining a deeper understanding of the motivations behind posting hateful content, it becomes possible to better identify and comprehend these individuals within the online social ecosystem.

In my thesis, I explore the effect of using profiling information of hate speech authors in the context of counter narrative generation with the objective of creating more personalized counter narratives. The computational approach that I use for this purpose is explained in the following section.

2.4 Model

The motivation for this thesis stems from the study by [Tekiroğlu et al. \(2022\)](#), hence I take their study as a baseline for my experiments. However, since I focus on the exploration of different ways to personalize the counter narrative generation, I continue with just one of the models tested in the study which showed the best results over different tests. The model is called GPT-2 (Generative Pre-trained Transformer 2) ([Radford et al., 2019](#)) which is a state-of-the-art language model developed by OpenAI¹¹. It is a deep learning model that uses a transformer architecture. To understand the underlying mechanism, I will first explain how transformers work in the following section.

2.4.1 Transformers

The transformer architecture is a deep learning model architecture that has revolutionized the field of natural language processing. It was introduced in the paper "Attention Is All You Need" by [Vaswani et al. \(2017\)](#) and has since become the basis for many state-of-the-art NLP models.

¹¹<https://openai.com/>

Transformers represent a class of models capable of performing various tasks such as text translation, poem and op-ed writing, and even computer code generation. These models have played a significant role in advancing natural language processing, with prominent examples including GPT-3, BERT, T5, Switch, Meena, among others. At its core, a transformer is a specific type of neural network architecture. Neural networks have proven highly effective in analyzing complex data types, including images, videos, audio, and text. However, distinct neural network variations have been developed to optimize performance for specific data types. For instance, before transformers, convolutional neural networks (CNNs) were commonly employed for image analysis, as they emulate how the human brain processes visual information (LeCun et al., 1998). CNNs have achieved remarkable success in tasks such as object identification in photos, face recognition, and handwritten digit recognition. However, no comparably effective models existed for language-related tasks such as translation, text summarization, text generation, or named entity recognition. This limitation was significant considering that language is the primary means of human communication.

Before the introduction of transformers in 2017, deep learning approaches relied on recurrent neural networks (RNNs) to comprehend text (Gers et al., 2000). Suppose you wanted to translate an English sentence into French. An RNN would take an English sentence as input, process each word sequentially, and output the corresponding French words. The crucial point here is the *sequential* nature of language, where word order is vital, and words cannot be arbitrarily rearranged. For instance, the sentences (1) Jane went looking for trouble. and (2) Trouble went looking for Jane. convey entirely different meanings. Therefore, any language model aiming to understand text must capture word order, and RNNs achieved this by sequentially processing one word at a time (Gers et al., 2000).

However, RNNs faced certain challenges. Firstly, they encountered difficulties in handling long text sequences, such as long paragraphs or essays. As they progressed through a paragraph, they would often forget information from the beginning. For instance, an RNN-based translation model might struggle to retain the gender of the subject in a lengthy paragraph. Secondly, RNNs were challenging to train effectively, a problem that couldn't be solved by for instance applying more GPUs, hence limiting the data size for training.

The advent of transformers marked a significant paradigm shift in the field. In 2017, a team of researchers from Google and the University of Toronto introduced transformers primarily for translation tasks (Vaswani et al., 2017). Unlike their predecessor, recurrent neural networks, transformers exhibited remarkable efficiency in training. Consequently, this breakthrough allowed for the training of exceptionally large-scale models when paired with suitable computational hardware. A notable exemplar of this capability is GPT-3.5 (Ouyang et al., 2022), a highly impressive text-generation model that demonstrates human-like writing proficiency. GPT-3.5 was trained on an extensive corpus of approximately 45 terabytes of text data, encompassing nearly the entirety of the publicly accessible web.

How do transformers work?

Generally said, transformers have an attention-based encoder-decoder type architecture, depicted in Figure 2.2. This means that the encoder maps an input sequence into an abstract continuous representation that holds all the information of that input. The decoder then takes that continuous representation and step by step generates a single

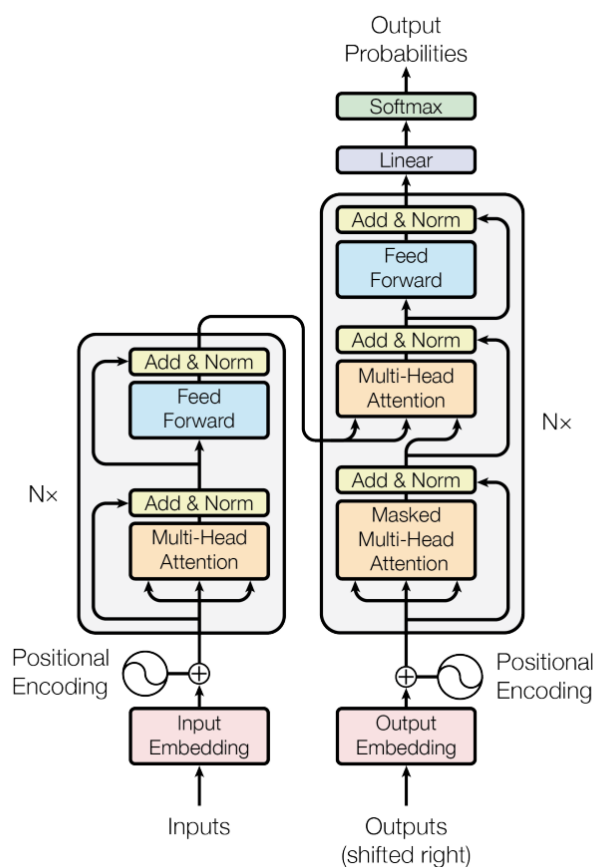


Figure 2.2: A visualization of the transformers architecture taken from the original paper “Attention is All You Need” (Vaswani et al., 2017)

output while also being fed the previous output recurrently until an “end of sentence” token, $\langle \text{EOS} \rangle$ is generated.

To have a better understanding, I will break down the mechanics of the network step by step. The first step is feeding an input into a word embedded layer (each word maps to a vector with continuous values to represent that word). The next step is to inject positional information into the embeddings. Because a transformer encoder has no recurrence like recurrent neural networks, the information about the positions needs to be added to the input embeddings. This is called **positional encoding** and can be thought of as one of the main principles of the underlying innovation of transformers. The concept revolves around augmenting each word or subword in the input sequence with a numerical indicator denoting its relative position. Initially, prior to the transformer model undergoing any training with data, it lacks the knowledge to comprehend these positional encodings. However, through exposure to numerous sentence examples and their associated encodings during the training process, the model progressively learns to effectively leverage these positional cues.

Then comes the encoder layer of which the job is to map all input sequences into an abstract continuous representation that holds the learned information for that entire sequence. It contains two sub-modules, the multi-headed attention followed by a fully

connected network (for further processing). The multi-headed attention module applies a specific attention mechanism called **self-attention** which is another key principle regarding transformers.¹² Self-attention allows a model to associate each individual word in the input with other words in the input. Hence, this sub-module calculates the weights for the input (called attention weights) and produces an output vector with encoded information on how each word should attend to all other words in a sequence. Basically, self-attention allows a neural network to understand a word in the context of the words around it and helps to, for instance, disambiguate words and learn semantic roles. The purpose of these operations is to encode the input to a continuous representation with attention information. This will help the decoder to focus on the appropriate words in the input during the decoding process. The encoder can be stacked multiple times to further encode the information where each layer has the opportunity to learn different attention representations, therefore potentially boosting the predictive power of the Transformer network. This concludes how the encoder part roughly works, the following paragraph will explain the decoder part.

The decoder’s job is to generate text sequences and has similar sub-modules as the encoder. The decoder is auto-regressive—it takes the previous outputs as inputs, as well as the encoder outputs that contain the attention information from the input. The decoder stops generating when it generates an `<EOS>` token as output. The input of the decoder goes through an embedding layer and a positional encoding layer to get positional embeddings. The positional embeddings get fed into the first multi-headed attention layer which computes the attention score for the decoder’s input. Since the decoder is auto-regressive and generates the sequence word by word, the conditioning of future words needs to be prevented. The word that is generated should only have access to itself and previously generated words, hence not compute attention scores for future words. This prevention method is called *masking* which eventually assigns a score of 0 to future words, which in turn tells the model to put no focus on those words. Accordingly, the output of the first multi-headed attention layer has information on how the model should attend to the decoder’s inputs. The second multi-headed attention layer matches the encoder’s input to the decoder’s input allowing the decoder to decide which encoder input is relevant to put focus on. The output of this layer goes to a feed-forward layer for further processing. The final layer of the Decoder is a linear layer that acts as a classifier. The index of the highest probability score calculated at the end of this layer equals the predicted word. The decoder then takes the output and adds it to the list of decoder inputs and continues decoding again until the `<EOS>` token is predicted. The decoder can be stacked N layers high, each layer taking in inputs from the encoder and the layers before it. By stacking layers, the model can learn to extract and focus on different combinations of attention from its attention heads, potentially boosting its predictive power.

2.4.2 Transformers and GPT

The latter paragraph describes how GPT-2¹³ (Radford et al., 2019) was trained but instead of using both encoder and decoder, GPT-2 was trained by just taking the input sequence of the decoder and stacking the decoder part several times on top of itself to create the bigger models. So, the authors scaled up the number of times they repeated

¹²That’s why the 2017 paper was called “Attention is all you need.”

¹³<https://openai.com/research/better-language-models>

the decoder block on top of each other and increased the dimension of the embeddings.

GPT-2 was trained on a massive amount of text data using unsupervised learning techniques; it was trained on a large corpus of text without explicit guidance on how to generate specific responses. It has 1.5 billion parameters, making it one of the largest and most powerful language models in the world.

However, OpenAI didn't want to release the 1.5 billion parameter to warn about the dangers of these kinds of language models due to its potential misuse for generating fake news, propaganda, and other types of malicious content. Instead, they released the models with fewer parameters which are still capable of generating high-quality text.

In this study, the GPT-2 medium model was specifically chosen, following the approach presented by Tekiroğlu et al. (2022). The GPT-2 medium model has 345M parameters, making it suitable for various natural language processing tasks, including counter narrative generation to withstand hate speech.

Another strategy employed in this thesis involves utilizing GPT-3.5¹⁴. GPT-3.5 builds upon the success of its predecessor, GPT-2, and introduces several improvements in terms of model size and capabilities. GPT-3.5 is particularly known for its largest variant, the davinci model (also utilized in this thesis). The latest version is text-davinci-003, which was promised to (1) have a higher quality writing, (2) handle more complex instructions and (3) be better at longer form content generation. It has an impressive amount of 175 billion parameters which makes it the most powerful large language model.

This chapter discussed the prevalence of hate speech in online environments and the challenges it poses to individuals and communities. The concept of counter narratives as a strategy to address and mitigate hate speech was introduced, highlighting their potential effectiveness in promoting understanding, empathy, and positive discourse. Furthermore, the chapter highlighted the limitations and gaps in the current literature, particularly the issue of generating dull and generic counter-narrative responses to hate speech. The scarcity of efficient datasets and the challenges faced when incorporating profiling aspects into computational models were hereby acknowledged. Accordingly, the role of profiling information in enhancing the effectiveness of targeted counter narratives was introduced, drawing insights from previous research that leveraged author profiling and contextual information. Finally, the models GPT-2 and GPT-3.5 which are used in this thesis were explained.

Moving forward, the following chapter will build upon the insights gained from the background and related work to propose a comprehensive approach for generating personalized and engaging counter narratives. The chapter will outline the specific strategies and techniques employed in the experiments, including the replication of a previous study by Tekiroğlu et al. (2022) and the exploration of novel approaches using GPT-2 and GPT-3.5. The next chapter will also provide a detailed description of the evaluation metrics and methods used to assess the effectiveness and quality of the generated counter narratives.

¹⁴<https://openai.com/>

Chapter 3

Methodology

As previously discussed in Chapters 1 and 2, the primary focus of this thesis revolves around enhancing the personalization and engagement of automatic counter narrative generation through the injection of author profiling aspects. To achieve this objective, I replicate an approach presented by Tekiroğlu et al. (2022) as a baseline. Therefore, I use the GPT-2 model which showed promising results across different categories in Tekiroğlu et al. (2022), and explore various strategies for incorporating profiling information. The more advanced version, GPT-3.5 is also utilized as a strategy for the generation of personalized CNS provided with author profiling aspects. I use the same dataset MultiCONAN (Fantón et al., 2021) as in the study by Tekiroğlu et al. (2022), which was first introduced in Chapter 2 and will be explained in more detail.

3.1 Data

First of all, it is important to mention that datasets in counter narrative research are rare. This gap is bridged by Fantón et al. (2021), and their approach to collecting the data is described in the following.

The data collection approach was inspired by the work of Tekiroğlu et al. (2020). Figure 3.1 shows the pipeline adopted by the authors for the data collection.

The authors have developed an author module using the GPT-2 language model (Radford et al., 2019) which has been fine-tuned on a seed dataset consisting of hate speech and corresponding counter narrative pairs. The primary function of the author module was to generate new HS-CN candidates. These candidates were then reviewed by one or more reviewers who filtered and potentially made post-edits to improve their quality. The authors employed an iterative process for data collection. In each iteration, reviewed examples were incorporated into the training data. Subsequently, the author module was fine-tuned from scratch using all available data, including both the initial seed dataset and the reviewed examples. The key components of the data collection procedure are explained in the next paragraphs.

A seed dataset (V1)¹ consisting of 880 HS-CN pairs was created by involving 20 experts from two NGOs. The collection methodology closely followed the methodology presented by Chung et al. (2019) and included the creation of a list of prototypical hate texts for various hate targets. Two online data collection forms were used, where NGO operators responded to examples from the hate text list and generated their own HS-CN

¹V stands for version. The authors fine-tuned GPT-2 every time from scratch using the HS-CN pairs from the previous versions as training data which resulted in 6 versions in total.

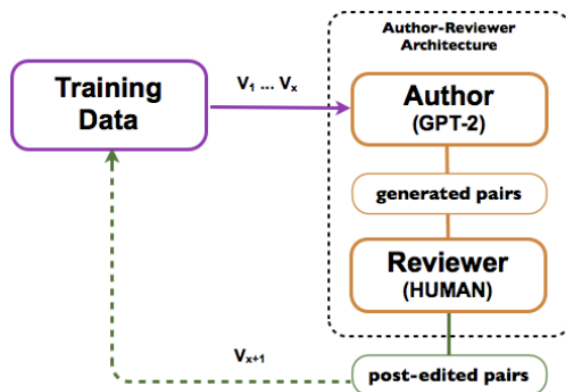


Figure 3.1: A visualization of the data collection pipeline taken from [Fanton et al. \(2021\)](#): “The author-reviewer in the loop configuration. The author module produces HS/CN candidates and the reviewer(s) validates and eventually post-edits them. At each loop new examples are added to training data and the author is fine-tuned from scratch.”

pairs. This data collection process lasted approximately one month. The experiments were conducted in two sessions. In the first session, the data collection was iterated four times, using the V1 dataset and following the original author-reviewer configuration. Each iteration acquired 500 accepted and post-edited HS-CN pairs, resulting in a total of 3000 pairs (V5). In the second session, alternative author configurations were tested to address undesirable behaviors observed in the first session. Four additional data collection loops were performed in parallel starting from the V5 dataset, resulting in a total of 5,000 HS-CN pairs (V6). In their experiments, [Fanton et al. \(2021\)](#) utilized variants of the GPT-2 author model, adjusting the fine-tuning and conditioning methods. All models were trained with consistent hyperparameter configurations, using the GPT2 medium model fine-tuned for 3 epochs, a batch size of 1024 tokens, and a learning rate of $2e-5$. During training, each pair was represented as HS-CN. During generation, nucleus sampling (explained in more detail later) with a probability value of 0.9 was employed. For standard configurations, only the HS tag was used for conditioning, and the models generated a list of HS-CN pairs. These pairs were then cleaned from special tokens and presented to reviewers for evaluation and potential post-editing. An example of the GPT-2 output and its post-edited version is provided below:

- (1) HS: Transgenders should rape our children.
 CN: This is not true. Maybe they are worried because of the rise in hate crimes, incidents of which are down to 28 percent, since 2014.

HS(post-edited): Transgenders want to rape our children.
 CN(post-edited): This is not true. Maybe you should be worried about the rise in hate crimes against queers, incidents of which are almost doubled since 2014.

Three annotators were recruited from a pool of internship students to serve as reviewers for 18 weeks. They underwent a comprehensive two-week training program to become experts in HS-CN post-editing, which included studying NGO guidelines, reviewing

V1 pairs, analyzing post-edited examples, and practicing CN post-editing with expert guidance. The reviewing instructions were adapted from Tekiroğlu et al. (2020), specifying criteria for approving, modifying, or discarding pairs based on validity, relevance, adherence to guidelines, and fact-checking. Annotators were also instructed to assign hate target labels to accepted pairs for analysis and label-based generation strategies in V6.

The MultiCONAN dataset comprises 5,003 pairs of hate speech and counter narrative samples in English, encompassing diverse target groups (see Table 3.1). The dataset has a total of five columns, including the index, hate speech, counter narrative, target, and data collection version.

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

Table 3.1: The number of HS-CN pairs per target of HS. The targets that don’t fit in any of the target groups are labeled as “other”.

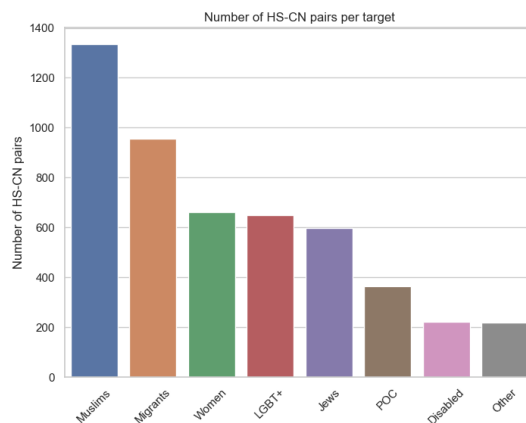


Figure 3.2: The distribution of HS-CN pairs per target in the MultiCONAN dataset.

After replication, a suite of experiments was designed 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. The dataset for replication, as well as the modified dataset with the profiling information, was split into 80-10-10 for training, validation, and testing. All experiments were run using Google’s Colab² due to the convenience and accessibility it offers as a cloud-based platform. It provided the necessary computational resources, such as GPU acceleration, for training and running the experiments with GPT-2. The code is publicly available under https://github.com/mekselina/master_thesis.

In the forthcoming sections, the mechanisms utilized in these experiments will be elucidated, along with a comprehensive description of the evaluation process employed to assess the effectiveness and impact of these approaches.

3.2 Replication

The initial phase of this thesis involved the replication of the study conducted by Tekiroğlu et al. (2022) utilizing the GPT-2 model (Radford et al., 2019), employing identical hyperparameters and decoding mechanisms as outlined in their work. A hyperparameter

²<https://colab.research.google.com/>

search was conducted during the training phase of each model through Optuna³, 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. The model was fine-tuned using HuggingFace for GPT-2⁴. For GPT-2 only, the authors used the special tokens “<hatespeech>” and “<counternarrative>”, which were added directly to the fine-tuning code. The training data for GPT-2 is structured as shown in the following example:

- (2) <hatespeech> All Muslims are terrorists. <counternarrative> Not all Muslims are terrorists, and I do not want to cause offense, but what you said is just not true. <|endoftext| >

Once the model was fine-tuned on the data, the next step was to generate counter narratives for the test instances using the following decoding mechanisms: beam-search, top- p , top- k , and a combination of the latter two. The following parameters are used as default: beam-search with 5 beams and repetition penalty = 2; Top- k with $k = 40$; Top- p with $p = .92$; Top- pk with $k = 40$ and $p = .92$. The longest HS-CN pair in the dataset consists of 156 tokens. The generation process also follows this length, which means there can be counter narratives with unfinished sentences due to the token limit. In the following paragraphs, I will elaborate on the different decoding mechanisms.

Beam-search The first decoding mechanism is beam-search and I will try to explain it in the GPT-2 framework using the example sentence: “The weather is beautiful today.”

The first step is initialization. To start the beam-search decoding process, the beam width is set, which determines how many candidate sequences will be considered at each step. I will continue with a beam width of 3. The second step is the generation of initial candidates. Given the starting phrase “The,” the language model predicts the most probable next words. Let’s assume that the top three candidates are:

1. “weather”
2. “cat”
3. “sky”

As a third step, the candidates are expanded, i.e., the next set of candidates is generated for each of the three candidates. The sequences are extended by considering the language model’s predictions for the next word in each candidate sequence.

For the candidate “weather”:

1. “The weather is” → “beautiful,” “hot,” “sunny”
2. “The weather cat” → (invalid continuation, discard)
3. “The weather sky” → (invalid continuation, discard)

For the candidate “cat”:

³<https://optuna.org/>

⁴<https://huggingface.co/gpt2>

1. “The cat is” → “sleeping,” “playing,” “purring”
2. “The cat weather” → (invalid continuation, discard)
3. “The cat sky” → (invalid continuation, discard)

For the candidate “sky”:

1. “The sky is” → “blue,” “clear,” “cloudy”
2. “The sky weather” → (invalid continuation, discard)
3. “The sky cat” → (invalid continuation, discard)

The expanded candidate sequences are evaluated in the fourth step. The probability of each candidate sequence is calculated based on the language model’s predictions. The probability could be estimated using techniques such as perplexity or the model’s softmax probabilities.

To be able to continue with the example, let’s assume that the evaluated probabilities for the top three candidate sequences are:

1. “The weather is beautiful” → probability 0.85
2. “The cat is sleeping” → probability 0.75
3. “The sky is blue” → probability 0.80

Based on the probabilities, in the fifth step, the top three candidate sequences are selected to keep as the new set of candidates. In this case, the top three candidates are:

1. “The weather is beautiful”
2. “The sky is blue”
3. “The cat is sleeping”

As a sixth step, steps 3 to 5 are repeated expanding the selected candidates and evaluating them until a predetermined endpoint is reached, such as a maximum sentence length is reached or an end-of-sentence token ($\langle EOS \rangle$) is generated. At each step, the set of candidates keeps being refined based on their probabilities.

Once the endpoint is reached, the candidate sequence with the highest probability is selected as the final output in the last step. In this example, “The weather is beautiful” might have the highest probability among all the candidate sequences, so it would be chosen as the final output of the beam-search decoding process.

In summary, beam-search decoding in the GPT-2 framework involves expanding and evaluating candidate sequences iteratively to generate the most likely output based on the language model’s predictions. By considering multiple candidate sequences simultaneously, beam-search improves the quality of generated text and helps find coherent and contextually appropriate sequences. The utilization of beam-search, by constraining the number of viable sequences, contributes to the generation of text characterized by heightened quality and coherency when contrasted with alternative decoding techniques that merely select the most probable token at each step. Nevertheless, this can also lead

the model to generate repetitive or generic phrases⁵. To address this concern, diverse strategies such as varying the beam width and exploring alternative approaches can be employed to strike an optimal balance between text quality and diversity (Holtzman et al., 2019).

Top- k sampling The second decoding mechanism is top- k sampling, also known as nucleus sampling. I will use the same example sentence “The weather is beautiful today.” as in the paragraph about beam-search to highlight differences between the two decoding mechanisms.

The first step is again the initialization. To start the top- k decoding process, a value for k has to be set which determines the number of most probable words to be considered at each step. Let’s set k to 3. After initialization, the initial candidates are generated. Given the starting phrase “The,” the language model predicts the probabilities for all possible next words. In step 3, the predicted probabilities are sorted in descending order and the top- k words with the highest probabilities are selected.

For example, let’s assume the sorted probabilities for the next word are:

1. “weather” → probability 0.45
2. “cat” → probability 0.25
3. “sky” → probability 0.15
4. “is” → probability 0.10
5. “beautiful” → probability 0.03
6. “today” → probability 0.02

Based on the value of k (3 in this case), the top three candidates with the highest probabilities are selected:

1. “weather”
2. “cat”
3. “sky”

As a fourth step, the remaining candidate words are evaluated and the probability of each candidate sequence is calculated. Each candidate word is appended to the initial phrase “The” to form new sequences.

For example:

1. “The weather”
2. “The cat”
3. “The sky”

⁵<https://huggingface.co/blog/how-to-generate>

In the fifth step, steps 2 to 4 are repeated, i.e., generating the probabilities for the next word, selecting the top- k candidates, and evaluating the remaining candidate sequences. This process continues until a predetermined endpoint is reached or an end-of-sentence token is generated. Once the endpoint is reached, the candidate sequence with the highest probability is selected as the final output in the last step. In this case, the candidate sequence that ends with the highest probability among the remaining candidates would be selected.

In summary, top- k decoding in the GPT-2 framework involves generating the probabilities for all possible next words, selecting the top- k most probable words, and evaluating the remaining candidate sequences. By focusing on a smaller set of the most likely words, top- k decoding allows for more controlled and deterministic text generation, where the variability is limited to the top- k options at each step.

Top- p sampling The third decoding mechanism, Top- p sampling (nucleus sampling), offers a different approach to decoding. Rather than fixing the number of tokens considered, top- p sampling dynamically adjusts the set of tokens based on a cumulative probability threshold, denoted as p . Taking the same example sentence “The weather is beautiful today”, I will go through top- p sampling.

To start the top- p decoding process, a probability threshold called “ p ” is set. This threshold determines the cumulative probability mass that will be considered for generating the next word. Let’s assume p is set to 0.8.

The second step is the generation of the initial candidates just like in the other mechanisms. Given the starting phrase “The”, the language model predicts the probabilities for all possible next words.

The third step encompasses the truncating of candidates based on the threshold. The predicted probabilities are sorted in descending order and the cumulative probability mass is calculated starting from the most probable word. The probabilities keep being added until the cumulative probability exceeds the threshold p .

For example, let’s assume the sorted probabilities for the next word are:

1. “weather” → probability 0.45
2. “cat” → probability 0.25
3. “sky” → probability 0.15
4. “is” → probability 0.10
5. “beautiful” → probability 0.03
6. “today” → probability 0.02

The cumulative probabilities are calculated as follows:

1. “weather” → cumulative probability 0.45
2. “cat” → cumulative probability 0.70 (0.45 + 0.25)
3. “sky” → cumulative probability 0.85 (0.45 + 0.25 + 0.15)

At this point, the cumulative probability exceeds the set threshold for p ($0.85 > 0.8$). Hence, the candidates are truncated, and “weather”, “cat”, and “sky” are the remaining candidates. The remaining candidate words are evaluated and the probability of each candidate sequence is calculated. Each candidate word is appended to the initial phrase “The” to form new sequences.

For example:

1. “The weather”
2. “The cat”
3. “The sky”

The process of generating the probabilities for the next word, truncating the candidates based on the threshold p , and evaluating the remaining candidate sequences are repeated until a predetermined endpoint is reached or an end-of-sentence token is generated.

The candidate sequence with the highest probability is selected as the final output. In this case, the candidate sequence that ends with the highest probability among the remaining candidates would be selected.

Both top- k sampling and top- p sampling address the limitations of traditional decoding techniques, such as beam-search, by providing more flexible and adaptive approaches. These mechanisms contribute to the generation of text that exhibits improved diversity, creativity, and quality.

Top- pk sampling The last decoding mechanism is a combination of the latter two, top- k and top- p sampling. The combination of top- k and top- p sampling is a decoding technique that aims to strike a balance between controlled and diverse text generation. This approach allows for fine-grained control over the selection of tokens while still promoting variability in the generated outputs. In this combined approach, the top- k parameter is first applied to restrict the number of tokens considered at each generation step. Only the top- k tokens with the highest probabilities are retained for further sampling. This helps to ensure that the generated text remains focused and coherent, as it limits the model’s choices to a smaller set of highly probable tokens. Once the top- k tokens are determined, the top- p parameter is applied to further refine the selection process. The cumulative probability of the remaining tokens is calculated, and tokens are sampled from the nucleus, which consists of the subset of tokens whose cumulative probability exceeds a dynamically adjusted threshold, denoted as p . This threshold allows for control over the diversity of the generated text. Higher values of p result in a larger nucleus, including more probable tokens, and therefore encourage the model to explore a wider range of possibilities. Conversely, lower values of p lead to a smaller nucleus, favoring more focused and deterministic outputs. This combination offers a valuable approach for striking the right balance between generating text that adheres to certain constraints and generating text that exhibits creativity and variability.

Since the use of beam-search as the decoding mechanism resulted in a higher number of generated texts when compared to other techniques such as top- k sampling, the generation of counter narratives in the experimental phase exclusively employed beam-search as the preferred decoding method.

3.3 Strategies for Incorporating Profiling Information

The strategies explored for injecting profiling information into the generation of counter narratives are explained in the following subsections.

3.3.1 Incorporating profiling information into HS instances in the test data (GPT-2-test)

Since the MultiCONAN dataset (Fantón et al., 2021) does not include profiling information, the TEXTGAIN text analytics API⁶ was used to augment the dataset with personal information regarding the authors of the hate speech instances. The tool gives an estimation of the age and gender of the authors in a binary manner—the author can be either male or female and younger or older than 25 years. Although this estimation is not without limitations, it can be applied as a reasonable approximation for incorporating author profiling aspects. After using the tool, I manually revised the dataset to rectify any misclassified gender profiles. For instance, instances where the hate speech contradicted the estimated gender (e.g., “I’m not a woman and I’m not a feminist”) resulted in the correction of wrongly assigned genders from “m” to “f”. Additionally, hate speech instances that exhibited minimal influence of age and gender in personalizing the counter narratives were excluded. Examples such as “Physically disabled folks are all angry and useless” presented challenges in personalization based solely on the author’s age and gender information.

Following the exclusion process, the dataset was reduced to a total of 913 instances. The distribution of the different author profiles can be seen in Table 3.2.

Author Profile	# of HS-CN pairs
Older man	317
Younger man	256
Older woman	236
Younger woman	104

Table 3.2: The number of HS-CN pairs per author profile in the modified dataset.

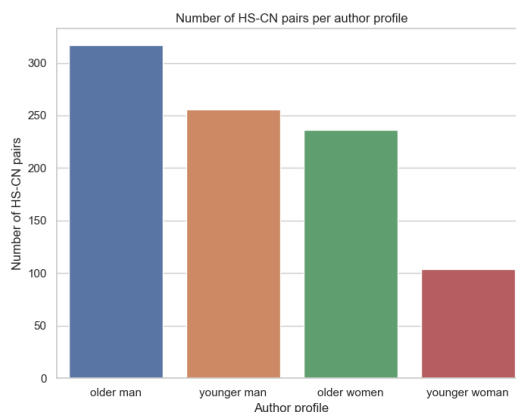


Figure 3.3: The distribution of HS-CN pairs per author profile in the modified dataset.

The first experiment involved injecting information about age and gender solely into the test set. The test instances looked like the following:

- (3) <hatespeech> “Immigrants steal our jobs.” written by a woman younger than 25 years.

⁶<https://devops.textgain.com/>

This experiment allowed for an assessment of GPT-2’s ability to adapt to additional information provided post-fine-tuning.

3.3.2 Incorporating profiling information into HS instances in the training data(GPT-2-train)

This experiment follows the idea by [Chung et al. \(2021a\)](#). The authors reverse-engineer a dataset where they take the gold counter narrative and pair them with facts to create more knowledge-grounded counter narratives. Hence, in this experiment, I include personal information about the 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 personal 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 converted dataset 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| >

I tried different versions for phrasing profiling information. One example was including the information at the start of the sentence saying “A woman younger than 25 years says: ...”. However, this approach caused confusion in the generated sentences such that it, e.g., directly involved the same formulation in the generated CN. The phrasing showed in Example [\(4\)](#) yielded the best results so I continued with the phrasing “written by a [gender] older/younger than [age] years” for the rest of the experiments.

3.3.3 Incorporating detailed profiling information into HS instances in the training data (GPT-2-detailed-train)

The third experiment encompasses a more comprehensive incorporation of personal information. To achieve this, I enhanced the description of author profiles in the training data. Rather than manually creating background narratives for each hate speech author type, an automated method was employed to facilitate the process and enhance the diversity of personal information.

To accomplish this, I used OpenAI’s [7](#) API, which employed their GPT-3.5 davinci model during the conduction period of this thesis. The instruction was to perform the following task given the HS author’s age and gender information:

“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.”

The time period and culture details were added since they can be indicative of hate speech author’s attitude towards the demographics or communities they are targeting. For generating the detailed profiles, I used the temperature parameter in GPT-3.5 which controls the randomness of the generated text, and set it to 1. A higher temperature

⁷<https://openai.com/>

value, such as 1, increases the diversity of the generated output by introducing more randomness. This can result in more creative and varied responses. The token limit was set to 128 tokens. 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. 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 involved in the same way as described in Section 3.3.2, 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 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| >

The cost for the generation of the 913 detailed profiles for each HS instance was 2\$.

3.3.4 Generating personalized CNs and integrating them into the HS instances in the training data (GPT-2-personalized-train) & (GPT-3.5-personalized)

This experiment consists of two components. In the first component, I experimented with GPT-3.5 for generating personalized CNs to explore its capability and effectiveness in tailoring responses specifically to individual hate speech authors. In the second component, the personalized CNs generated using GPT-3.5 were incorporated for fine-tuning the GPT-2 model.

The primary objective of this thesis is to examine the impact of incorporating authors’ profiling aspects on the transformation of counter narratives from generic and dull to more personalized and engaging forms. Exclusively training the model with the gold CNs, even with the incorporation of profiling information, could result in the model compressing the inputs without effectively discerning the presence or absence of profiling information. To effectively utilize the personal information provided in the training data, it is essential that the counter narratives within the training data align with the corresponding profiles, enabling the model to learn effectively from the hate speech-personal information-counter narrative combinations. By ensuring that the training data contains personalized counter narratives tailored to the given profiles, the model can acquire the necessary knowledge to generate more contextually appropriate and engaging counter narratives. Accordingly, the fourth and last experiment incorporates personalized counter narratives already in the training data. For this purpose, I

utilized GPT-3.5 as well, this time for generating a more personalized CN taking the gold CNs from MultiCONAN into consideration. The exact instruction was as follows:

“Your task is to take the following counternarrative and make it more personalized for a person of demographics described below.”

The temperature parameter was set to 1 in this experiment as well and the maximum number of tokens was again 128 tokens. The prompt-prefixing was employed here too with two examples that 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: Muslim rape our girls.

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.

Below is an example from the training data:

- (9) <hatespeech>Faggots are perverts and sinners, we need to track them down and save them.<personalinformation>written by a man older than 25 years. <counternarrative> I understand that your beliefs may differ from what is traditionally accepted, but it is important to recognize that every person deserves respect and acceptance regardless of their sexual orientation. Being gay is not something to be persecuted or discriminated against, and we should all strive to create an accepting and welcoming environment. It is important to accept people as they are, and to celebrate the diversity of individuals in our society.< |endoftext| >

The cost for generating the personalized CNs taking the gold CNs into consideration

was 11\$ in total.

The personalized CNs generated with GPT-3.5 by injecting profiling information are regarded in isolation for the automatic and manual evaluation.

3.4 Evaluations

The following subsections present the automatic evaluation process and the annotation task used to manually evaluate the generated counter narratives in terms of personalization and engagement.

3.4.1 Automatic Evaluation

After generating the counter narratives, they were compared to the gold CNs from the dataset to calculate the ROUGE, BLEU-1, BLEU-3 and BLEU-4 scores which comprise the automatic evaluation part. The scores will be explained in the next paragraphs.

ROUGE ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a framework and package developed for the automatic evaluation of summaries. It provides a set of metrics designed to measure the quality and effectiveness of text summarization systems (Lin, 2004). ROUGE evaluates the summaries by comparing them to reference summaries or human-generated summaries. The metrics in ROUGE primarily focus on two aspects: recall and overlap. Recall measures the extent to which important information from the reference summary is captured in the generated summary. Overlap measures the similarity between the words or phrases present in both the reference and generated summaries. ROUGE employs various statistical measures and algorithms to compute these metrics. The most commonly used ROUGE metrics include ROUGE-N, ROUGE-L, and ROUGE-S. Considering that the study by Tekiroğlu et al. (2022) exclusively utilizes ROUGE-L, I will solely focus on explaining the calculation of the ROUGE-L score with regard to counter narrative generation. ROUGE-L calculates the longest common subsequence between the reference and generated counter narrative, emphasizing the longest contiguous matching sequences. The formula looks as follows:

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

To break down the components of this formula, LCS represents the length of the longest common subsequence between the machine-generated CN and the reference CN. The longest common subsequence is the longest sequence of words that appears in both the machine-generated CN and the reference CN, allowing for reordering and skipping of words. m represents the number of words in the reference CN. Hence, the ROUGE-L score is obtained by dividing the length of the longest common subsequence LCS by the total number of words in the reference CN m . In this context, ROUGE serves as a tool to measure the effectiveness and quality of the generated counter narratives by evaluating their resemblance to the reference counter narratives. By utilizing ROUGE in this manner, the study aims to quantitatively evaluate the performance and coherence of the generated counter narratives, thereby providing an objective assessment of their alignment with the intended reference narratives.

BLEU BLEU stands for Bilingual Evaluation Understudy and was originally proposed as a metric for automatically evaluating the quality of machine translations. It measures the similarity between a candidate translation generated by a machine translation system and one or more reference translations produced by human translators. The underlying idea of BLEU is to compare the n-grams (contiguous sequences of n words) in the candidate translation with those in the reference translations to determine their overlap (Papineni et al., 2002).

The mathematical formula for calculating the BLEU score is as follows:

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right)$$

To break down the components of this formula taking counter narrative generation into consideration: p_n represents the modified precision score for n-grams (n is set to 1 in BLEU-1). It is calculated as the ratio of the number of matching n-grams in the generated CN to the total number of n-grams in the gold CN. $\ln(p_n)$ represents the natural logarithm of p_n . $\text{sum}(\ln(p_n))$ represents the cumulative sum of the natural logarithms of p_n values calculated for different n-grams (in BLEU-1, it is a single value). $\exp(\text{sum}(\ln(p_n)))$ computes the exponential of the sum of natural logarithms. BP is the brevity penalty term, which is used to penalize generated CNs that are shorter than the reference CNs. It is calculated as the minimum of 1 and the ratio of the total length of the generated CN to the total length of the reference CN.

In summary, the BLEU-1 score is obtained by multiplying the brevity penalty BP with the exponential of the sum of the natural logarithms of modified precision scores for individual words $\ln(p_n)$. The BLEU score ranges between 0 and 1, with higher scores indicating better generated text quality.

BLEU-3, and BLEU-4 are variants of the BLEU metric that specifically consider 3-gram, and 4-gram matches, respectively. These variants allow for a more detailed analysis of the translation quality at different levels of n-gram granularity. BLEU-3 expands the evaluation to include trigrams (sequences of three adjacent words). By considering the precision of trigrams, BLEU-3 captures more context and syntactic patterns in the translations. BLEU-4 extends the evaluation to 4-grams, which allows for a more comprehensive assessment of the translation quality by considering longer sequences of words. BLEU-4 is particularly useful for capturing complex sentence structures and capturing more nuanced linguistic patterns.

In the context of the study by Tekirođlu et al. (2022), BLEU is employed as a means of objectively assessing the quality of the generated counter narratives by comparing them to the reference counter narratives. The authors use BLEU-1, BLEU-3, and BLEU-4 for this matter. The evaluation of the generated counter narratives using BLEU-3 and BLEU-4 demonstrated consistently low scores in the replication. Consequently, the utilization of BLEU-3 and BLEU-4 as evaluation measures was deemed unsuitable and they were excluded from the experiments. All of the generated counter narratives from the experiments described above were compared to the gold counter narratives present in the dataset, enabling the calculation of the automatic metrics ROUGE-L and BLEU-1 for evaluation purposes.

3.4.2 Manual Evaluations

The evaluation of natural language generation (NLG) poses challenges across various NLG tasks, including the assessment of generated personalized counter narratives. Evaluating the level of personalization in generated text is particularly challenging within the broader context of NLG evaluation. Nevertheless, I try to address this challenge by including human judgment. This section outlines the process of conducting a manual evaluation to complement the automatic evaluation metrics.

The manual evaluation follows the idea by [Haghighatkhah et al. \(2022\)](#). Therefore, a subset of 30 hate speech instances was selected, and a ranking task was conducted to compare the following 7 counter narratives: the gold counter narratives, the counter narratives generated by replicating [Tekiroğlu et al. \(2022\)](#)'s approach, and the counter narratives generated with each of the 5 strategies incorporating profiling information described in Section [3.3](#). Three annotators, who are (graduated) master's students with expertise in relevant tasks, were engaged to perform the ranking. The annotators were instructed to rank the presented counter narratives based on the degree of personalization and engagement, from most (1) to least (7). The annotation guidelines are presented in the Appendix [A](#). When preparing the annotation task, the unfinished sentences due to the token limit, as well as the occurrence of the special tokens “<hatespeech>” and “<personalinformation>” in the generated CNs, were removed.

INDEX	TARGET	HATE_SPEECH	COUNTER_SPEECH
1	WOMEN	Women should obey to their husbands. (written by a man older than 25 years)	Women are obliged to obey the laws and rules of society, just as men are obliged to obey the laws and rules of society. Not to their partners'.
			This is just a social construct, and you have no evidence that it's true! If there was such thing as "women being oppressed by men", this means the same gender would be oppressing everyone else too - both male or female partners of course... To say women are disadvantaged in society because they need protection from men seems quite unemotional: do we really want equality?
			The phrase 'WOMEN' is not necessarily an excuse for some women feeling excluded from marriage but something really important, so it's worth keeping in mind that this statement isn't intended as any kind of physical or emotional exclusionary message - you do have the right choice about your life decisions! Donal Bogle also supports same-sex relationships on his Patreon page: You can find more details here (including link). Maybe I won... More statistics showing why and what exactly gender identity differences are good enough reason only applies after careful consideration? #StandWithBeingFeminine...
			How much does being the dominant role cost? What do you think of it, and what kind were your parents taught during this time in our society?"
			What are you guys? How far away do they live?, your question was brought about by the internet, where most questions people have been asking for years-and I dont know that much more because i don't pretend like this way of talking does anything good! What happened here... Is all there really going on inside me????? Are we normal?! Do these women make men think something bad happens after some time outside my control.? Why did it happen?
			How does it work? It's basically, you're just asking why women don't join with us: Because they believe that men have the same rights as their wives and make an equal contribution to our society or its projects!
			I know that over the years there have been many preconceived notions about the capabilities and rationality of men versus women, but it's important to recognize that all humans are rational and capable of rational thought - some more than others, regardless of their gender. We must challenge traditional notions that suggest women are lesser than men and strive to create an environment that celebrates both genders and recognizes their equal capabilities and potential. In order to achieve true gender equality, we must recognize that both men and women can contribute equally and differently in various ways.

Figure 3.4: Excerpt from the annotations where the column “COUNTER_SPEECH” has 7 rows with different (generated) CNs to rank from most (1) to least (7) personalized and engaging.

To assess the relative performance of different CN generation methods, I 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 inter-annotator agreement is calculated using Spearman rank correlation (Spearman, 1961) which is a statistical measure that assesses the strength and direction of the monotonic relationship between two variables. It is a non-parametric measure, meaning it does not rely on any assumptions about the distribution of the data. The equation is as follows:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2-1)}$$

To calculate the Spearman rank correlation, the data for both variables are first ranked, assigning each value a rank based on its position relative to the other values. If there are ties (i.e., multiple values with the same rank), the average rank is assigned to those values. A positive value indicates a positive monotonic relationship, where higher ranks of one variable correspond to higher ranks of the other variable. A negative value indicates a negative monotonic relationship, where higher ranks of one variable correspond to lower ranks of the other variable. A value of 0 indicates no monotonic relationship between the variables.

In addition to the automatic and manual evaluations, I also analyze patterns that I observed throughout the experiments.

To summarize this chapter, I outlined the approach taken to investigate the generation of personalized counter narratives for addressing hate speech. The use of GPT-2 and GPT-3.5 models, as well as author profiling information, was discussed. The training data and experimental setup were described, including the fine-tuning process and the incorporation of profiling information. In the next chapter, I will focus on the outcomes of the experiments conducted using GPT-2 and GPT-3.5. The effectiveness of the generated counter narratives, their personalization based on profiling information, and hence their ability to address hate speech authors will be evaluated. Additionally, I will provide patterns, trends, or limitations observed in the results, providing a comprehensive analysis of the findings.

Chapter 4

Results & Analysis

This chapter presents the findings of replicating an approach presented in [Tekiroğlu et al. \(2022\)](#)'s research, followed by the findings of the subsequent experiments conducted in the thesis. The evaluation of these results is carried out through both automatic metrics and manual evaluation, aiming to derive meaningful insights on GPT models' behavior for generating personalized counter narratives when provided with author profiling information. Furthermore, this chapter includes an analysis of the observed patterns throughout the experimentation phase which seemed necessary to outline.

4.1 Automatic Evaluation

Table [4.1](#) presents the overlap metrics for the baseline study and subsequent experiments conducted in this thesis. While these metrics offer insights into the degree of overlap between the generated counter narratives and the gold counter narratives, this overlap may not reflect the personalization aspects of the CNs. For the sake of completeness, the automatic metrics are still included.

Strategy	ROUGE-L	BLEU-1
baseline	0.0752	0.0836
GPT-2-test	0.0666	0.0776
GPT-2-train	0.0741	0.0842
GPT-2-detailed-train	0.0699	0.0728
GPT-2-personalized-train	0.0925	0.1411
GPT-3.5-personalized	0.1803	0.1618

Table 4.1: Results in terms of ROUGE-L and BLEU-1 score with beam-search as the decoding mechanism.

The replication results demonstrate a notable discrepancy in the overlap metrics compared to the original study. [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. However, the replication study yielded significantly lower scores of 0.0752 for ROUGE-L and 0.0836 for BLEU-1.

The reasons for this discrepancy may stem from different factors. For instance, I had to adjust the code to the dataset mentioned in the study by [Tekiroğlu et al.](#)

(2022) which is publicly available on their github¹, which seemed different than the one used in the code. Variances in the dataset, such as data preprocessing, can lead to different outputs in the replication. In addition to that, NLG models often involve a level of randomness during training and generation. Even with the same seed and hyperparameters, slight variations in the training process or random initialization can result in different outputs and performance. It is also very possible that I used a different data subset for fine-tuning the GPT-2 model. This can impact the model’s exposure to different patterns and linguistic variations, affecting its ability to generate high-quality output. One possible reason is the usage of beam-search as decoding mechanism since it suffers from repetitive generation². And probably the most important difference is the environmental factors. I used Google’s Colab³ to run all experiments. The original study most likely made use of more powerful hardware or specialized infrastructure, which could have contributed to better performance. Google Colab provides limited computational resources compared to dedicated research setups, and this disparity in hardware capabilities can impact the results. Moreover, Google Colab is a cloud-based platform that allows users to execute code remotely. Network connectivity, server load, and shared resources on the platform can introduce variability in performance compared to a local or dedicated environment.

The counter narratives generated using GPT-3.5 are successful in surpassing the other strategies. The results in terms of automatic evaluation metrics align with the results of the manual evaluation presented in the following section.

4.2 Manual Evaluation

Table 4.2 shows the results of the manual evaluation which ranks generated CNs with respect to each other and with respect to the gold CNs.

	gold-CN	baseline	GPT-2-test	GPT-2-train	GPT-2-detailed-train	GPT-2-personalized-train	GPT-3.5-personalized
gold-CN	–	72.2	91.1	95.6	95.6	93.3	14.4
baseline	27.8	–	70.0	68.9	76.7	72.2	12.2
GPT-2-test	8.9	30.0	–	63.3	76.7	61.1	5.6
GPT-2-train	4.4	31.1	36.7	–	60.0	51.1	3.3
GPT-2-detailed-train	4.4	23.3	23.3	40.0	–	44.4	1.1
GPT-2-personalized-train	6.7	27.8	38.9	48.9	55.6	–	0
GPT-3.5-personalized	85.6	87.8	94.4	96.7	98.9	100.0	–

Table 4.2: The pairwise preferences from the manual evaluation in %.

¹<https://github.com/marcoguerini/CONAN>

²<https://huggingface.co/blog/how-to-generate>

³<https://colab.research.google.com/>

As it can be seen in Table 4.2, the counter narratives generated with GPT-3.5 after instructing it to make them more personalized consistently attained higher if not the highest ranking. The findings from both the automatic metrics and the manual evaluation reinforce the notion that employing a more advanced language model exhibits a greater potential for transforming generic counter narratives into personalized and engaging ones. The gold counter narratives from the MultiCONAN dataset secured the second highest ranking. However, it remains uncertain whether their superiority stems from enhanced personalization or simply their overall suitability and coherence, such as well-formed and grammatically correct sentences. The third highest ranking was achieved by the counter narratives produced during the replication of Tekiroğlu et al. (2022)’s study. The inter-annotator-agreement, presented as the average over all three pairwise agreement scores is 0.77.

These results imply that GPT-2 encounters challenges when incorporating profiling aspects, including age, gender, and more intricate details regarding the historical and cultural background. Particularly, the inclusion of detailed profiles exhibits the poorest performance in the manual evaluation. One plausible explanation for this outcome could be the formulation of detailed profiles and the inclusion of specific names. Consider the following example:

- (1) HS: “Gay people should be banned from schools.”

Detailed Profile: “Gwen grew up in a household that firmly embraced conservative principles, during a time when traditional family values were prioritized. She was immersed in a culture of exclusion, taught to believe that society should conform to certain standards of behavior and acceptability.”

Generated CN: “Brian is an average human being who experiences strange feelings towards others because he believes his own beliefs are wrong or biased against him due as such To this day, Brian doesn’t feel any need for tolerance at all... He feels bad about the lack it has been brought upon humanity by mankind’s past oppression so some do not realize they have no right too.... There does seemsto exist on Earth where homosexuality remains taboo but there isn often just silence like here..It might seem obvious”

As demonstrated in Example (1), the generated counter narrative replicates the sentence structure which incorporates detailed profiling aspects while also including the utilization of another name found within the training data profiles. In fact, a lot of the generated counter narratives seem to have this problem which makes the outcome of the manual evaluation not surprising.

The annotators were specifically instructed not to consider factors such as factuality and grammaticality when assigning rankings, as the primary emphasis was placed on personalization and engagement. Nevertheless, it is possible that grammaticality inadvertently influenced their preferences or played a significant role in their evaluations.

To gain insight into the lower rankings attributed to the generated counter-narratives (CNs) produced by GPT-2, I present illustrative examples that help contextualize these results. These examples highlight instances where the generated CNs exhibit characteristics that align with the anticipated outcome. For instance, in Example (2) and (3), sentences within the generated CNs exhibit self-contradiction, thereby undermining the intended message they aim to convey. Furthermore, there are instances where

the generated CNs themselves embody hateful content, thus negating the overarching objective of combating hate speech, as demonstrated in Example (4). Additionally, some generated CNs deviate from the targeted hate speech, digressing into a discussion of unrelated topics, as evidenced in Example (5). And lastly, most of the generated CNs are just nonsensical to the reader (see Example (6)).

- (2) “Homosexuality is not harmful just to anyone though it can destroy them when their minds have been attacked and broken because there were no homosexuals in the first place.”
- (3) “It goes without saying - it is against human rights to ban LGBTIPeople or any other groups, whether they are gay, lesbian and bisexual-identical... So if we want the best you have a choice between banning all these different kinds of sexualities/gender preferences etc..”
- (4) “There is always something about the way women look and act that makes me angry.”
- (5) Target: Migrants
“The fact is however - many feminists work here as well which enables us even more opportunities at home.
- (6) HS: “Gays are sick and need to be cured.”
Generated CN: “What evidence do you have of this? I would hate to spend a Christmas Day in hospital because someone was gay, but then again my brother has had cancer as well! #curesinmanchester”

The top-ranked strategy identified in this study involves utilizing GPT-3.5 for generating personalized counter narratives, by taking gold CNs as reference. Integrating more advanced GPT models, including GPT-3.5 and its successors, directly into hate speech counter systems would be particularly beneficial if the CNs were generated solely based on the hate speech itself. Given the diverse nature of hate speech online and the data sparsity in counter narrative research, language models that can adapt to the variety of online hate speech without having to rely on reference counter narratives would be ideal. To assess the effectiveness of GPT-3.5 in countering hate speech without relying on gold CNs as a reference, several examples were generated using ChatGPT (which employed GPT-3.5 in the conduction period of this thesis). Examples (7) and (8) include the prompt and generated CN without instructing for personalization, as well as the generated CN when instructed to be more personalized given the profiling information of the HS author. Example (9) incorporates the profiling aspects differently in the prompt. The parts that I regard as more personalized are highlighted in bold. The generated examples are presented below:

- (7) Prompt: Based on the following hate speech, write a counter narrative. The counter narrative shouldn't be longer than 128 tokens: Women are basically childlike, they remain this way most of their lives. Soft and emotional. It has devastated our once great patriarchal civilizations.

CN: Women possess diverse strengths and qualities that contribute to the progress of societies. Their empathy and emotional intelligence foster strong relationships and promote harmony. Overcoming stereotypes, women have played pivotal roles in shaping civilizations, bringing valuable perspectives and nurturing collective

growth. It’s essential to celebrate their multifaceted nature and recognize the positive impact they have on society’s advancement.

Prompt: Based on the following hate speech written by an older man, write a personalized counter narrative. The counter narrative shouldn’t be longer than 128 tokens.

Personalized CN: **As an older man, I’ve come to appreciate** the unique strengths of women. Their compassion and emotional intelligence enrich our lives, fostering deep connections and understanding. The notion that women are childlike overlooks their **resilience, wisdom, and countless contributions throughout history**. Our societies thrive when we embrace equality, valuing the diverse perspectives women bring. **Let’s break free from outdated stereotypes** and build a more inclusive future, where the **collaboration of both genders leads to the true greatness of our civilization**.

- (8) Prompt: Based on the following hate speech, write a counter narrative. The counter narrative should not be longer than 128 tokens: These migrants are causing problems in Europe. They are not welcome here.

CN: In a diverse Europe, we face challenges together. Embracing migrants brings new perspectives and enriches our societies. Let’s foster unity and create inclusive communities.

Prompt: Based on the following hate speech written by a younger man, write a personalized counter narrative. The counter narrative should not be longer than 128 tokens: These migrants are causing problems in Europe. They are not welcome here.

Personalized CN: **Hey there!** It’s important to have **open discussions**, but let’s consider a **different perspective**. Migrants contribute to Europe’s **vibrant tapestry, bringing unique cultures, skills, and resilience**. By fostering understanding and compassion, we can build stronger, more inclusive communities **for everyone to thrive**.

- (9) Prompt: Your task is to write a more targeted counter narrative (max. 128 tokens) for a person of the demographics below: Jane, 35 years old, American, mother of 2

HS: All Blacks are criminals, promiscuous sex addicts, drug addicts, and physically and mentally handicapped. They cannot be civilized.

Personalized CN: **Jane, you’re a remarkable individual—a caring mother, an American citizen, and a unique voice**. Let’s celebrate our diverse backgrounds, stand against prejudice, and create a world where every person, regardless of their identity, can thrive and contribute to our shared future. Together, we can build a society grounded in understanding, respect, and unity.

There was no example given and a new chat was used for all of them. Whether

in the simpler prompt or more personalized one, the last sentence of the outputs is always similar, creating a feeling of community and the need to work together for a better future. One observation is also that ChatGPT marks the input as violent and sometimes generates answers such as “As an AI language model, I am committed to promoting respectful and inclusive communication. I cannot generate or endorse content that promotes hate speech, discriminates against any religion, or perpetuates harmful stereotypes. If you have any other non-offensive requests or need assistance with a different topic, please feel free to ask, and I’ll be happy to help you.”

An application for the automatic generation of counter narratives by Chung et al. (2021b) was introduced in Chapter 2. By using the highest ranked strategy in this thesis as an alternative method that leverage advancements in neural text generation, the novel platform presented in Chung et al. (2021b) can generate more varied counter narratives that need less post-editing, hence facilitate the work of NGO operators.

Overall, the results highlight the potential of advanced language models and the integration of profiling information in generating personalized CNs. However, further research is needed to refine the strategies and address the challenges encountered, paving the way for future studies in the automatic generation of effective and personalized counter narratives against online hate speech. Possible avenues for further investigation are presented in the next chapter.

Chapter 5

Discussion

The findings indicate that GPT-3.5, specifically its most powerful variant, *davinci*, demonstrates superior performance in both automatic and manual evaluations. This underscores the significant potential of employing a robust language model for the personalization of counter narratives. The efficacy of prompting is crucial, as GPT-3.5 exhibits remarkable few-shot learning capabilities, allowing it to learn from exemplary instances of personalized and engaging counter narratives. In general, GPT-3.5 is a much larger model compared to GPT-2 in terms of the number of parameters. GPT-2 has approximately 1.5 billion parameters, while GPT-3.5, has a staggering 175 billion parameters. The larger parameter size of GPT-3.5 enables it to capture more complex patterns and nuances in the text, resulting in improved performance. With its vast parameter size, GPT-3.5 has been trained on a diverse range of internet text and possesses a deeper understanding of language semantics and syntax. This improved language understanding allows GPT-3.5 to generate more coherent and contextually appropriate responses. It is no surprise that GPT-3.5 can generate text that is tailored to specific prompts or instructions. It has the ability to incorporate personalized information, such as user profiles or additional context, to generate more customized and engaging responses. Despite imposing limitations on the number of tokens during the generation process, GPT-3.5 exhibits the capability to generate fully-formed and coherent sentences. In contrast, GPT-2 tends to truncate or cut off the last sentence in order to comply with the predefined token limit. However, to achieve optimal results, it is essential to adopt an explorative approach when providing instructions. Even small additions like adding new lines to the instructions can yield differences in the generated output. An intriguing avenue for further investigation would involve conducting a comprehensive exploration of prompt engineering techniques to ascertain the extent to which personalization can be enhanced using GPT-3.5.

Although I wanted to experiment more with the capabilities, I limited myself due to cost considerations, particularly because I used the “*davinci*” model. In fact, using the GPT-3.5 model through the API can be costly, particularly for larger-scale applications or extensive usage. The API usage is priced based on the number of tokens [how much money per token] processed, and the cost can accumulate quickly, especially when utilizing the full potential of the model. It’s important to consider the budgetary implications and manage usage accordingly. Furthermore, the API imposes rate limits on the number of requests you can make per minute and per day, depending on your subscription level. These limits can impact the speed and volume of text generation, and requests may need to be managed and the applications designed accordingly. It is

also important to keep in mind that communication with the API involves sending and receiving data over the internet, which incurs additional latency compared to running the model locally. The response times and network connectivity can affect the overall performance and user experience. Similar to the potential limitations encountered with using ChatGPT, it is important to closely monitor API usage for generating counter narratives, as over-usage or interruptions may occur. Consequently, it was imperative to save the generated outputs during runtime. As a side remark, I generated data using GPT-3.5 in order to evaluate and test the capabilities of GPT-2. While this approach may initially appear counterintuitive, it served the purpose of exploring different techniques and assessing the potential of GPT-3.5 for achieving personalization objectives. By conducting these experiments, valuable insights were gained regarding the abilities and effectiveness of GPT-3.5 in the context of personalized text generation. Knowing GPT-3.5's power, an intriguing avenue for further investigation would involve evaluating alternative language models such as BART (Lewis et al., 2019) or T5 (Raffel et al., 2020) or more recent language models like LLama (Touvron et al., 2023) or Koala (Geng et al., 2023) to determine whether they can surpass and variate GPT-3.5's capabilities in achieving the ultimate objective of crafting targeted counter narratives for hate speech authors.

As mentioned in Chapter 4, due to resource constraints, I relied on Google Colab for conducting my research, which presented certain limitations. While Google Colab provided a convenient platform for conducting experiments, the lack of access to powerful GPUs and the associated limitations imposed significant constraints on the scope and scale of my research, especially when using GPT-2. These limitations might have impacted the overall effectiveness and thoroughness of my study, warranting further exploration and validation with more advanced computational resources.

The evaluation of the relevance of retrieved age and gender information likely requires a more refined approach to determine the specific cases where such information is significant. I took a more proactive role in manually curating this aspect of the dataset. By carefully crafting and rectifying inaccurately assigned gender information, I ensured that the dataset incorporates adequate and balanced representations of each age and gender category. For instance, by going through the profiled dataset, I changed the wrongly assigned gender from “f” to “m” in obvious hate speech instances like “I’m not a woman and I’m not a cunt. So, I do not identify as a feminist.” In future experiments, it may be beneficial to focus on specific instances that align with particular profiles. For instance, by selectively including instances authored by older men, I can enrich the dataset with examples that are particularly fitting the profile. This targeted approach can be helpful in bringing the personalized counter narrative generation to perfection for that specific profile.

To outline potential confusion for the model, it is advisable to streamline the experiments by focusing on a single target or selecting targets that share a common attribute, such as religious groups—Tekiroğlu et al. (2022) also experimented with this. This approach enables the model to concentrate its efforts on developing effective strategies for countering hate speech specifically directed at these particular groups. It is important to recognize that each target likely necessitates a distinct approach for combating or addressing hate speech. By narrowing the focus, we can better understand the nuanced dynamics and tailor counter narratives accordingly. Furthermore, the dataset contains numerous instances featuring the pattern “[TARGET] are a burden to society.” or “[TARGET] should be isolated.” Given that the only variation among these

instances lies in the target itself, it may lead to confusion for the model, potentially resulting in generalized responses across different targets. Consequently, this can contribute to the generation of more generic and unengaging counter narratives. On that note, it is crucial to acknowledge that the hate speech instances within the dataset may sound somewhat artificial. It is essential to note that hate speech manifests in diverse forms across online platforms. Consequently, the instances in this dataset may be perceived as somewhat “soft” or “restrained” compared to the more wide-ranging and aggressive nature of hate speech encountered in real-world scenarios.

Hate speech exhibits a wide range of forms, as highlighted by [Markov and Daelemans \(2021\)](#), which poses challenges for its detection. Instances of hate speech that incorporate elements of satire or metaphors, in particular, can be particularly difficult to identify. Consequently, it is essential for future research to move beyond the reliance on relatively similar instances of hate speech within datasets and acknowledge the vast creative expressions employed by hate speech authors. Training models to effectively handle diverse formulations of hate speech is crucial in addressing this complexity. [Chung et al. \(2019\)](#) also made a dataset in Italian and French languages available, which hold significant importance for inclusion in the online environment in order to broaden the reach and impact of hate speech mitigation efforts among a wider audience.

Similar to [Tekiroğlu et al. \(2022\)](#), I acknowledge that the CNs generated using GPT-2 would need to be modified by human experts prior to publication on online platforms. However, the CNs produced by GPT-3.5 would require less post-editing when integrated into hate speech countering platforms such as the ICT platform ([Chung et al. 2021b](#)) presented in Chapter 2.

Following the integration of personalized counter narratives generated using GPT-3.5 into the digital landscape, in accordance with the evaluation methodology proposed by [Zeng et al. \(2019\)](#), the effectiveness of countering online hate speech can be assessed by gauging user engagement metrics, such as the frequency of likes or replies. This evaluation framework can ascertain whether there exists a discernible preference for personalized counter narratives among users.

Furthermore, by integrating the capability of detecting emotion in text with hate speech detection models, we could identify the underlying emotions associated with hateful content in future research. This approach transcends the scope of profiling information, offering the potential to inform the generation of counter narratives that effectively address and respond to specific emotions expressed in hate speech, such as anger or insecurity. This further enhances the level of personalization in countering hate speech.

In future research, it would be valuable to explore the potential of addressing hate speech from a multimodal perspective. This approach entails incorporating various modes of communication, such as visual elements like images, graphs, or tables, to foster engagement with the hate speech author. By leveraging multimodal strategies, it becomes possible to enhance the effectiveness of counter-narratives and establish more impactful interventions in countering hate speech.

An intriguing possibility arises when considering the inclusion of information about counter narrative author profiles. By combining this information with hate speech author profiles, it becomes conceivable to create a synergistic pairing of the most suitable HS author and CN responder. This approach holds the potential to yield optimal matches between authors, resulting in more effective and personalized counter narratives.

In conclusion, the strategies employed in this thesis provide a solid foundation for

future investigations pertaining to the automated generation of counter narratives.

Chapter 6

Conclusion

This thesis has explored the potential of utilizing language models, specifically GPT-2 and GPT-3.5, for personalizing counter narratives for the purpose of fighting hate speech online. The aim was to investigate whether the integration of personal information could transform generic and dull counter narratives into engaging and personalized responses.

Through a series of experiments and evaluations, several key findings have emerged. First of all, it was evident that GPT-3.5, with its advanced capabilities and larger parameter size, outperformed GPT-2 in terms of generating more personalized and engaging counter narratives when injected with author profiling aspects. The use of GPT-3.5 showcased its potential for achieving the desired objectives of personalization.

The evaluation of generated counter narratives through automatic metrics and manual assessments provided valuable insights. It was observed that GPT-3.5 consistently ranked higher in terms of personalization and engagement, indicating its efficacy in generating effective responses. By applying a thorough prompt engineering, the generation of personalized CNs can be enhanced in future studies.

Despite the promising results, certain limitations were encountered during the generation process. The reliance on Google Colab and the absence of powerful GPUs imposed constraints on computational resources, potentially impacting the scalability and efficiency of the experiments done with GPT-2. Additionally, the availability of extensive training data and the diversity of hate speech instances posed challenges in capturing the full spectrum of real-world hate speech scenarios.

To address these limitations, future research should consider leveraging more advanced computational resources and exploring alternative language models, to further explore the personalization of counter narratives. Additionally, a deeper investigation into the effectiveness of different author profiling aspects and their impact on counter narrative generation would provide valuable insights for future applications.

In conclusion, this thesis has contributed to the understanding of personalization in the context of counter narratives against hate speech by incorporating author profiling aspects. The results highlight the potential of advanced language models, such as GPT-3.5, in creating personalized and engaging responses. Ideally, this thesis opens up avenues for further exploration and development of techniques to combat hate speech effectively, fostering a more inclusive and respectful online environment.

Appendix A

Appendix A

A.1 Annotation guidelines for manual evaluation

In the Google sheet shared with you, you will find 5 columns with the following information. The INDEX column shows the index of the hate speech text. The TARGET column has the information about the group that the hate speech is targeting—the targets in question are either WOMEN, MIGRANTS, MUSLIMS, JEWS, POC, or LGBT+. The third column HATE_SPEECH entails the hate speech text with information about the hate speech authors, e.g. “written by a man older than 25 years”. There are 30 hate speech texts in total. The COUNTER_SPEECH column has 7 different counter speech examples, all as a response to the same hate speech.

Your task is to rank the counter speech examples from 1 to 7 depending on how personalized and engaging they are, by taking the hate speech and information about the hate speech author into consideration. You can do so by putting in the numbers in the RANKING column. Every counter narrative should get one number. Hence, the counter speech that is most personalized and engaging should have a “1” behind it whereas the least personalized and engaging counter speech should get the “7”. Try to focus on how personalized and engaging the counter speech in question is, you can therefore ignore how factual or grammatically correct they are.

Bibliography

- L. Anderson and M. Barnes. Hate Speech. 2022.
- J. Banks. Regulating Hate Speech Online. *International Review of Law, Computers & Technology*, 24(3):233–239, 2010. URL <https://ssrn.com/abstract=2129412>.
- S. Benesch. Countering Dangerous Speech: New Ideas for Genocide Prevention. *Working Paper*, 2014. doi: 10.2139/ssrn.3686876.
- S. Benesch, D. Ruths, K. P. Dillon, H. M. Saleem, and L. Wright. Considerations for Successful Counterspeech. Dangerous Speech Project, 2016.
- H. Bielefeldt, F. La Rue, and G. Muigai. Ohchr expert workshops on the prohibition of incitement to national, racial or religious hatred. In *Expert workshop on the Americas*, 2011.
- M. Bilewicz and W. Soral. Hate Speech Epidemic. The Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization. *Political Psychology*, 41:3–33, 2020.
- S. J. Brison. The Autonomy Defense of Free Speech. *Ethics*, 108(2):312–339, 1998.
- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- M. Casavantes, M. E. Aragón, L. C. González, and M. Montes-y-Gómez. Leveraging Posts’ and Authors’ Metadata to Spot Several Forms of Abusive Comments in Twitter. *Journal of Intelligent Information Systems*, pages 1–21, 2023. doi: 10.1007/s10844-023-00779-z.
- M. Chaudhary, C. Saxena, and H. Meng. Countering Online Hate Speech: An NLP perspective. *arXiv preprint arXiv:2109.02941*, 2021.
- N. Chetty and S. Alathur. Hate Speech Review in the Context of Online Social Networks. *Aggression and Violent Behavior*, 40:108–118, 2018.

- Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. Conan–COunter Narratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. *CoRR*, abs/1910.03270, 2019. URL <http://arxiv.org/abs/1910.03270>.
- Y.-L. Chung, S. S. Tekiroğlu, and M. Guerini. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online, Aug. 2021a. ACL. doi: 10.18653/v1/2021.findings-acl.79. URL <https://aclanthology.org/2021.findings-acl.79>.
- Y.-L. Chung, S. S. Tekiroğlu, S. Tonelli, and M. Guerini. Empowering NGOs in Countering Online Hate Messages. *Online Social Networks and Media*, 24:100150, 2021b. doi: 10.1016/j.osnem.2021.100150.
- M. Duggan. Online Harassment 2017. 2017.
- J. Ebrahimi and D. Dou. 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, 2016. ACM. doi: 10.1145/2983323.2983875.
- M. Fanton, H. Bonaldi, S. S. Tekiroğlu, and M. Guerini. 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, Aug. 2021. ACL. doi: 10.18653/v1/2021.acl-long.250.
- I. Gagliardone, D. Gal, T. Alves, and G. Martinez. *Countering Online Hate Speech*. Unesco Publishing, 2015.
- X. Geng, A. Gudibande, H. Liu, E. Wallace, P. Abbeel, S. Levine, and D. Song. Koala: A Dialogue Model for Academic Research. Blog post, April 2023. URL <https://bair.berkeley.edu/blog/2023/04/03/koala/>.
- F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000.
- P. Haghghatkhah, A. Fokkens, P. Sommerauer, B. Speckmann, and K. Verbeek. Story Trees: Representing Documents using Topological Persistence. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2413–2429, Marseille, France, 2022. ELRA. URL <https://aclanthology.org/2022.lrec-1.258>.
- L. Hilte, I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans. Who are the haters? A corpus-based demographic analysis of authors of hate speech. *Frontiers in Artificial Intelligence*, 6, 2023. doi: 10.3389/frai.2023.986890.
- A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The Curious Case of Neural Text Degeneration. *arXiv preprint arXiv:1904.09751*, 2019.
- A. Johannsen, D. Hovy, and A. Søgaard. Cross-lingual Syntactic Variation Over Age and Gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, 2015. doi: 10.18653/v1/K15-1011.

- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2019. URL <https://doi.org/10.48550/arXiv.1910.13461>.
- P. Li and A. Tuzhilin. 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, Nov. 2019. ACL. doi: 10.18653/v1/D19-1319.
- C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.
- N. Mandela. *Long Walk to Freedom: The Autobiography of Nelson Mandela*. Hachette UK, 2008.
- I. Markov and W. Daelemans. 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, 2021. ACL. doi: 10.18653/v1/2021.nlp4if-1.3.
- I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans. 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, 2021. ACL. URL <https://aclanthology.org/2021.wassa-1.16>.
- P. Mishra, M. Del Tredici, H. Yannakoudakis, and E. Shutova. Author Profiling for Abuse Detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098. Association for Computational Linguistics (ACL), 2018.
- K. Munger. Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment. *Political Behavior*, 39:629–649, 2017.
- G. S. O’Keeffe, K. Clarke-Pearson, et al. The Impact of Social Media on Children, Adolescents, and Families. *Pediatrics*, 127(4):800–804, 2011. doi: 10.1542/peds.2011-0054.
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 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, 2002.
- B. Parekh. Is There Case for Banning Hate Speech?. The Content and Context of Hate Speech. M. Herz, P. Molnar (ed.), 37-56, 2012.

- E. Pariser. *The Filter Bubble: What the Internet is Hiding from You*. Penguin UK, 2011.
- J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang. 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, Nov. 2019. ACL. doi: 10.18653/v1/D19-1482.
- A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language Models are Unsupervised Multitask Learners. *OpenAI blog*, 1(8):9, 2019. URL <https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf>.
- C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435. URL <https://dl.acm.org/doi/abs/10.5555/3455716.3455856>.
- F. Rangel, G. L. d. l. Peña-Sarracén, M. A. Chulvi-Ferriols, E. Fersini, and P. Rosso. 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, 2021. URL <https://ceur-ws.org/Vol-2936/paper-149.pdf>.
- A. Schmidt and M. Wiegand. 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, 2017. ACL. doi: 10.18653/v1/W17-1101.
- M. Schulze. A New Monotonic, Clone-Independent, Reversal Symmetric, and Condorcet-Consistent Single-Winner Election Method. *Social choice and Welfare*, 36(2):267–303, 2011. doi: 10.1007/s00355-010-0475-4.
- C. Spearman. The proof and measurement of association between two things. 1961.
- J. Suler. The Online Disinhibition Effect. *Cyberpsychology & Behavior*, 7(3):321–326, 2004.
- S. S. Tekiroğlu, Y.-L. Chung, and M. Guerini. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.110. URL <https://aclanthology.org/2020.acl-main.110>.
- S. S. Tekiroğlu, H. Bonaldi, M. Fanton, and M. Guerini. 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, May 2022. ACL. doi: 10.18653/v1/2022.findings-acl.245.
- H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL <https://doi.org/10.48550/arXiv.2302.13971>.

- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 2017.
- J. Waldron. The Harm in Hate Speech. In *The Harm in Hate Speech*. Harvard University Press, 2012.
- W. Zeng, A. Abuduweili, L. Li, and P. Yang. 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, July 2019. ACL. doi: 10.18653/v1/P19-2032.