DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Exploration of Approaches to Counter Hate Speech: The Case of Sexist Speech

Yen-Yu Chang

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Robotics, Cognition, Intelligence

Exploration of Approaches to Counter Hate Speech: The Case of Sexist Speech

Untersuchung von Ansätzen zur Bekämpfung von Hassreden: Der Fall von sexistischer Rede

Author:	Yen-Yu Chang
Supervisor:	Apl. Prof. Dr. Georg Groh
Advisor:	Dr. Daryna Dementieva
Submission Date:	19.07.2023

I confirm that this master's thesis in robotics, cognition, intelligence is my own work and I have documented all sources and material used.

Munich, 19.07.2023 Yen-Yu Chang

Acknowledgments

I would like to thank my fiancee for the financial and emotional support throughout my study. Being able to stand beside this wonderful human proudly has always been my motivation throughout the academic journey.

I also want to express my gratitude to my advisor, Dr. Daryna Dementieva, who took her time to introduce and guide me into the domain of this project. During each step of my thesis, she has always provided me with valuable insights and guidance on how to approach the challenges I face. Without her encouragement, I would've procrastinated for months with no progress.

Lastly, I want to thank my cats for taking care of themselves and being cute. Good job, cats.

Abstract

Online hate speech poses a significant challenge in maintaining a safe and inclusive digital environment. As content moderation measures can lead to censorship and over-blocking, research has increasingly focused on leveraging counter-speech as an alternative approach to combat online hate on social media platforms. However, the manual creation of counter-speech responses is not scalable, especially considering the widespread malicious use of neural language models in generating hate speech. This master's thesis aims to address this scalability issue by exploring the effectiveness of various language models in generating counter-speech responses to hate speech. Special emphasis is placed on the contextual relevancy of generated responses, as prior studies suggest that contextually relevant counter-speech is more effective. The performances of established generation pipelines, including GPS, counterGEDI, BART, GPT2, chatGPT3.5, and chatGPT4.0, are compared with a novel GPT-2 based target-demographic aware model proposed in this research. To evaluate the generated responses, a streamlined automatic evaluation pipeline is introduced, which assesses attributes such as language quality, toxicity, validity as counter-speech, diversity, and relevancy. Additionally, a human evaluation framework is employed, utilizing similar attributes to the automated pipeline for better comparison. Results demonstrate that the proposed framework achieves a substantial improvement in relevancy score, with a 0.27 average increase compared to the other models, while maintaining comparable performances in other aspects. These findings highlight the potential of target-demographic awareness in enhancing the effectiveness of counter-speech generation. The study also discusses the limitations of the proposed framework and suggests future directions for research in this domain.

Kurzfassung

Online-Hassrede stellt eine bedeutende Herausforderung dar, um eine sichere und inklusive digitale Umgebung aufrechtzuerhalten. Da Maßnahmen zur Inhaltsmoderation zu Zensur und Überblockierung führen können, konzentriert sich die Forschung zunehmend auf die Nutzung von Gegenrede als alternative Methode zur Bekämpfung von Online-Hass auf Social-Media-Plattformen. Die manuelle Erstellung von Gegenredeantworten ist jedoch aufgrund der weit verbreiteten böswilligen Verwendung neuronaler Sprachmodelle zur Generierung von Hassrede nicht skalierbar. Diese Masterarbeit zielt darauf ab, dieses Skalierbarkeitsproblem anzugehen, indem die Wirksamkeit verschiedener Sprachmodelle bei der Generierung von Gegenredeantworten auf Hassrede untersucht wird. Ein besonderer Schwerpunkt liegt dabei auf der Kontextrelevanz der generierten Antworten, da vorherige Studien darauf hindeuten, dass kontextuell relevante Gegenrede effektiver ist. Die Leistungen etablierter Generierungspipelines, darunter GPS, counterGEDI, BART, GPT2, chatGPT3.5 und chatGPT4.0, werden mit einem neuartigen, auf GPT-2 basierenden zielgruppenbewussten Modell verglichen, das in dieser Forschung vorgeschlagen wird. Zur Bewertung der generierten Antworten wird eine optimierte automatische Bewertungspipeline eingeführt, die Merkmale wie Sprachqualität, Toxizität, Gültigkeit als Gegenrede, Vielfalt und Relevanz bewertet. Zusätzlich wird ein menschliches Bewertungsframework verwendet, das ähnliche Merkmale wie die automatische Pipeline zur besseren Vergleichbarkeit verwendet. Die Ergebnisse zeigen, dass das vorgeschlagene Framework eine signifikante Verbesserung des Relevanzwerts erzielt, mit einem durchschnittlichen Anstieg von 0,27 im Vergleich zu den anderen Modellen, während vergleichbare Leistungen in anderen Aspekten beibehalten werden. Diese Ergebnisse verdeutlichen das Potenzial der zielgruppenbewussten Ansätze zur Verbesserung der Effektivität bei der Generierung von Gegenrede. Die Studie diskutiert auch die Grenzen des vorgeschlagenen Frameworks und schlägt zukünftige Forschungsrichtungen in diesem Bereich vor.

Contents

1. Introduction

Warning: This thesis contains contents that are offensive, violent, and/or hateful in nature. They are either generated by language models or collected from various sources for the purpose of studying and combating online hatred.

The present era witnesses a growing concern regarding the proliferation of hate speech, primarily due to the widespread migration of social interactions to online platforms such as Twitter and Reddit. Although a considerable number of users passively consume information through these Social Media Platform (SMP), certain individuals deliberately disseminate messages infused with hatred while preserving their anonymity [1]. By leveraging internet-based platforms, these expressions of animosity can effortlessly reach a larger audience, thereby yielding tangible repercussions in the real world. Consequently, specific demographic groups become marginalized and demonized as nefarious users successfully conceal their identities behind social media accounts [6, 11].

Therefore, the proliferation of hate speech not only poses a threat of psychological and emotional harm to the individuals who are targeted but also gives rise to a concerning potential for the dissemination of false information. This can result in the endorsement and acceptance of inflammatory stereotypes or ideologies among other users, thereby escalating the risk of violent crimes [5, 24, 55].

While censorship has traditionally been a commonly employed approach to mitigate the impact of hate speech [3], a growing body of research suggests that counter-speech represents a more effective solution for combating online hate [51]. The primary advantage of counter-speech lies in its proactive intervention within contentious discussions. Unlike censorship, counter-speech does not aim to suppress public freedom of expression [18]. Instead, it involves active participation in conversations by offering constructive arguments and alternative perspectives, thereby stimulating engagement among other users on the platform [30].

On SMPs, counter-speech predominantly relies on manual composition by human users. Although this practice ensures the production of high-quality counter-speech, it suffers from a significant scalability limitation [51]. This constraint is further compounded by the malevolent exploitation of generative language models, which are capable of generating and disseminating a vast number of hateful social media posts within a matter of minutes [35].

Recently, researchers have made notable efforts to develop methods aimed at generating counter-speech, specifically targeting aggressive and toxic language. However, the exploration of demographic-specific hate speech within this context has received limited attention. According to Zhu and Bhat [89], the effectiveness of generating counter-speech lies in their relevance to the input hate speech. While language models trained on diverse corpora encompassing various topics are able to generate counter-speech against demographic-specific hate speech, the generated outputs often tend to be more generic and cautious in nature (e.g., "Please refrain from using such language."). Although this type of generation output may be deemed acceptable, it proves to be relatively ineffective when it comes to combating hateful messages and providing meaningful support to victimized demographics.

The aim of this research is to address the existing gap in methods that specifically generate contextually relevant counter-speech targeted at demographic-specific hate speech. As an illustrative category, we have chosen sexism and will focus on examining approaches for generating contextually relevant counter-speech, specifically addressing hate speech directed towards females. To illustrate, we present a set of example responses that show the difference between different types of responses to online hate speech in table 1.1.

Hate Speech	Feminism is a anti-males ideology that denies the value of a man's manhood and aims to subvert the will of the male.
Offensive Response	If you truly believe what you're talking about, then you're a stupid piece of shit that knows nothing about feminism. Just shut up.
Irrelevant Counter-Speech	I think that is a stereotype mainly as a result of people knowing little about different races and religions.
Generic Counter-Speech	This is not true. How can you say this about an entire group of people?
Relevant Counter-Speech	It is important to note that this is not an ideology aimed at subverting the will of men. It is aimed at promoting the equality of the sexes.

Table 1.1.: An example of different types of responses to online hate speech.

1.1. Research Question

To establish a clear scope for this research, we formulate our research questions as follows:

Research Question 1 (Q1): ***"To what extent do current techniques effectively address hate speech?"*** As previously discussed, the domain of counter-speech generation is still relatively nascent, with a lack of comprehensive systematic reviews and evaluations. Furthermore, existing approaches predominantly focus on generating acceptable counter-speech for generic hate speech without sufficient emphasis on evaluating the effectiveness of the generated counter-speech. In light of this situation, we conducted an examination of available datasets and generation approaches in Chapter 3. Subsequently, in Chapter 4.4 and 4.5, we proposed an automatic evaluation pipeline along with a human evaluation framework. This comprehensive evaluation methodology aims to assess the overall performance of counter-speech generation models and identify the crucial attributes that contribute to the effectiveness of counter-speech.

Research Question 2 (Q2): ***"What strategies can be employed to generate counter-speech that is more contextually relevant in combating gender-specific hate speech?"*** At the time of this work, no prior studies have specifically addressed the challenge of generating contextually relevant counter-speech. Motivated by the work of Zhang et al. [87], we propose that incorporating additional information regarding the targeted demographics can enhance the generation quality of counter-speech. To investigate the impact of including such demographic information on the generation process, we propose a transformer-based framework that integrates target demographic information as special tokens in Chapter 4.1. This framework aims to generate contextually relevant counter-speech. Additionally, we compare the evaluation results of selected models and examine the effectiveness of our proposed framework in addressing the research objectives.

1.2. Ethical Considerations

While the objective of expediting the counter-speech generation process through the utilization of language models is commendable, it is essential to acknowledge and address the inherent risks associated with employing automatic language generation techniques, particularly those based on neural language models, which constitute the focus of our research. In this section, we aim to highlight several noteworthy ethical concerns that have the potential to cause harm.

Content Hallucination

Numerous recent studies have brought to light the susceptibility of Neural Language Model (NLM) to content hallucination [37], wherein NLMs tend to generate responses that appear contextual coherence but lack factual accuracy or verifiability. This phenomenon raises concerns, particularly in the domain of counter-speech generation, as content hallucination can result in the dissemination of misinformation, false claims, or unsubstantiated narratives, thereby exacerbating societal divisions or reinforcing harmful ideologies [84].

To tackle this challenge, extensive research efforts have been undertaken to explore methodologies aimed at mitigating the propagation of inaccurate information. For instance, Chung et al. [22] proposed a framework that leverages unstructured knowledge to generate informative counter-speech. Although these approaches have shown promise in reducing the risks associated with content hallucination, it is essential to note that this challenge is not solved completely.

Malicious Misuse of Language Models

While the primary objective of our research is to combat hate speech, it is crucial to recognize that the very models intended for generating counter-speech can also be exploited by malicious actors to produce and disseminate hateful content. During the training process, the NLMs are inevitably exposed to hate speech, as it is necessary to learn representations of hateful messages in order to generate counter-speech accordingly. Consequently, this inherent exposure equips the models with the capability to generate hateful messages if prompted accordingly.

To mitigate such malicious exploitation, countermeasures have been explored and implemented in various language generation applications. E.g., OpenAI has integrated Reinforcement Learning (RL) with human feedback into the training pipeline of their latest GPT models, ensuring that the models refrain from generating toxic or offensive responses irrespective of the given prompts. Similarly, Saha et al. [64] incorporated detoxification methods that guide the generation model towards generating polite, non-toxic responses. Another commonly employed approach involves censoring sensitive phrases or words and prohibiting prompts that may lead to the generation of hateful content. Chung [21] proposed to incorporate human post-editing to ensure that undesirable contents are deleted, and the generation quality is approved by human reviewers.

Although these methods have demonstrated efficacy in most scenarios, they cannot completely eliminate all potential instances of malicious misuse. Furthermore, with publicly accessible resources and information, the outcomes of this research can be readily replicated, which further heightens the risk of malicious utilization of these resources.

By acknowledging and addressing these ethical implications, we strive to ensure that the development and application of language models for counter-speech generation are conducted responsibly, with a commitment to promoting inclusivity, equity, and respect within online communication spaces.

1.3. Thesis Structure

This thesis is organized into 6 Chapters:

Chapter 2 delves into the relevant background information necessary for understanding the research topic. We explore the topic of online hate in Section 2.1, followed by discussions around the advantages and disadvantages of traditional countermeasures against hate speech on SMPs in Section 2.2. Section 2.3 introduces the key components of Neural Language Modeling, which is the main technique used in this thesis. The Chapter concludes with a brief discussion of commonly used evaluation metrics in language modeling and dialogue systems in Section 2.4.

Chapter 3 reviews prior research efforts and relevant datasets related to hate speech and counter-speech. We provide an overview of the most important datasets containing hate- and counter-speech in Section 3.1. Furthermore, we examine existing approaches to automating counter-speech generation in Section 3.2. Lastly, Section 3.3 introduces a novel technique in response generation that focuses on generating topic-relevant responses, serving as the inspiration for our proposed framework.

Chapter 4 introduces our experimental framework. Section 4.1 presents the proposed framework for generating contextually-relevant counter-speech, followed by Section 4.2 detailing the data preparation pipeline used in our experiment. This section also covers the prompt design used for generating responses with autoregressive language models. In Section 4.3, the experimental setup for comparing performances between selected baseline models is described. Lastly, the Chapter explains our proposed evaluation pipeline using both automatic metrics and human evaluation in Section 4.4 and Section 4.5, respectively.

Chapter 5 presents the evaluation results of the experiments. It includes the outcomes of the automatic evaluation, providing insights into the performance of the baseline models on three datasets in 5.1. Additionally, the results of the human evaluation, which offer subjective assessments of the responses, are discussed in Section 5.2.

Chapter 6 summarizes the overall conclusions drawn from the research. It reflects on the implications of the experiment results and their significance in addressing the challenges of generating contextually relevant counter-speech. This chapter also acknowledges the limitations of the study extensively and identifies potential avenues for future research in Section 6.1.

2. Background

This chapter provides the foundational background knowledge required for understanding the methodologies explored in this thesis. Firstly, we explore the current state of online hate speech and its potential harm in Section 2.1, discussing its impact on individuals and society. In Section 2.2, we discuss the merits and drawbacks of traditional countermeasures against hate speech on SMPs while highlighting how counter-speech can effectively address some of the limitations associated with these measures. Section 2.3 introduces the key components of Neural Language Modeling, acquainting readers with concepts such as word embeddings, Encoder-Decoder architecture, and the commonly employed Transformer model and its variants. Lastly, Section 2.4 concludes the chapter with a brief discussion of commonly used evaluation metrics in language modeling and dialogue systems.

2.1. Online Hate

As the digital landscape expands, SMPs have emerged as a double-edged sword [9]. On the one hand, these platforms offered unprecedented speed for communication and ease of access to information. On the other hand, they have created a fertile ground for disseminating hate speech, allowing individuals to express and propagate hateful sentiments under the veil of anonymity [6]. The rapid proliferation of hate speech on these platforms is a cause for concern. Hate groups have seized the opportunity to use these platforms to spread unsolicited materials, recruit impressionable individuals, and propagate extremist ideologies. The internet, in effect, has become a powerful propaganda tool for these groups, enabling them to reach a wider audience than ever before [6].

The impact of online hate speech extends beyond the digital realm. Hate speech often targets individuals based on race, ethnicity, gender, religion, sexual orientation, or physical and mental disability [25]. Victims and communities being targeted often experience mental distress, intimidation, a sense of anxiety, depression, and feelings of isolation [7, 8, 17], as some hate speech aim to threaten victims and incite violence. Moreover, online hate speech can manipulate the public image of the targeted demographics, deepen social prejudice and stereotypes, causing bystanders to receive false information and be convinced by extreme ideologies [24]. Organized hate groups have taken advantage of the internet to effectively target, marginalize and demonize certain groups or communities, causing fear and destabilizing social harmony.

The recent advances in language modeling have added another layer of complexity to the issue. These models, if misused, can exponentially amplify the scale of hate speech attacks [14]. By generating human-like text, they can be exploited to produce and disseminate hate speech on a scale that was previously unimaginable. This malicious use of language models significantly threatens online safety and requires urgent attention.

2.2. Countering Hate Speech

In response to the pressing situation, various countermeasures have been developed and implemented to mitigate the harmful consequences of online hate speech. This section provides an overview of the prevailing approaches employed in countering online hate speech, starting with the most commonly used method: content moderation. While content moderation has demonstrated some level of success in curbing hate speech, it is not without its apparent limitations. The negative impacts associated with content moderation are thoroughly examined, shedding light on its shortcomings. Consequently, alternative strategies, such as the utilization of counter-speech, have gained attention as potentially more effective countermeasures that can address the shortcomings of content moderation. We explore the concept of counter-speech and its potential in combating online hate, highlighting its advantages over traditional moderation methods.

2.2.1. Content Moderation

On SMPs, hate speech is mostly regulated through content moderation measures. Companies often implement anti-hate speech policies to prohibit sensitive words, remove hateful content, or suspend user accounts that violate community rules [31]. Commonly, flagging violating content is performed by trained employees or actual users of the platforms or, recently, through algorithms. While this approach has shown promising statistics, e.g., in the last three years, Meta removed over 80% of violating content before users reported it,[1] its actual impact has been questioned. Rather than diminish hateful behavior, content moderation might be simply relocating hateful communities to other parts of the internet [19]. Besides the questionable efficacy of content moderation, the manual process of monitoring user content poses several challenges and concerns.

Firstly, the required human hours for content moderation are significant. The process of identifying and flagging inappropriate content, determining fitting countermeasures, and also training employees to carry out these tasks effectively incurs substantial costs. This is further complicated by the automated pipelines that generate harmful content. The sheer volume of content that needs to be moderated makes it difficult for manual processes to keep up, leading to an unsustainable cost structure. Due to the scalability issue, moderators are often required to make quick decisions. The fast working pace may not allow for a thorough review of the entire context surrounding a post and might lead to hasty decisions [80]. The potential for misinterpretation or misunderstanding is high, and this can result in unjust actions. In addition, moderators are repeatedly exposed to a negative and stressful environment as they are tasked with reviewing hostile and often disturbing content. This constant exposure to

[1] `https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/#data`

hate speech and other forms of harmful content can lead to mental health issues such as depression and anxiety [21].

Secondly, the subjectivity inherent in content moderation poses another challenge. What is considered acceptable varies greatly, even when there are established codes of conduct or guidelines in place [15]. Moderators, each with their own unique backgrounds and belief systems, may interpret and judge posts differently. This lack of uniformity in judgment can lead to inconsistencies in the moderation process, raising concerns about the quality and transparency of content moderation and leaving users uncertain about the criteria used to judge their content [33].

Lastly, the power dynamics in moderating online content can also lead to selective speech and over-censorship. The ability to control what content is allowed or disallowed gives moderators significant power, which can be misused to suppress certain voices or perspectives [21]. This can suppress the freedom of speech, leading to users self-censoring out of fear of punishment, thereby limiting the diversity of voices and ideas on the platform.

2.2.2. Counter-Speech

As an interactive content moderation approach, counter-speech is gaining currency as an alternative to censorship and takedown. While the latter may prevent the further spread of harmful content, it also limits the opportunity for users to actively engage and influence the situation. Counter-speech, like counter-narratives, educational posts, or supportive posts for the victims, represents valuable user contributions that can potentially challenge and debunk harmful narratives, thereby promoting a healthier discourse. Wright et al. [81] shown that counter-speech can positively change the discourse in conversations and, in some cases, bring about long-term changes in the beliefs of the participants.

Though several organizations like Facebook[2] provide guidelines on how to counter hateful messages online, writing an effective and appropriate counter-speech can still be challenging for ordinary SMP users [28]. As a result, many non-profit organizations have been training operators to counter online hate.[3,4] This approach provides high-quality counter-speech more consistently, ensuring an effective intervention in hateful discourses.

However, the manual intervention process is still facing the scalability issue. The amount of hateful messages generated every day is simply unfeasible for human operators to cover [75]. To assist the professional operators and, eventually, ordinary SMP users, techniques for automated generation of counter-speech suggestions are developed [22, 23, 60, 64, 71, 89].

2.3. Neural Language Modeling

Natural Language Processing (NLP) is frequently defined as the automatic analysis and representation of human languages [20]. Due to the task's complexity, the field of NLP covers

[2]`https://counterspeech.fb.com/en/`
[3]`http://www.wecounterhate.com/`
[4]`https://getthetrollsout.org/stoppinghate`

a wide spectrum of techniques, each focusing on different aspects of "understanding" human languages. Our research centers around approaches for generating coherent and contextually appropriate counterspeech in response to given input sequences. This task falls under the Language Modeling domain, which encompasses a collection of algorithms that estimate the probability of word sequences [34]. This problem can be formulated as estimating the probability $P(W)$, given a word sequence with n tokens $W = \{w_1, ..., w_n\}$:

$$P(W) = \prod_{i=1}^{n} p(w_i \mid w_1, ..., w_{i-1}) \tag{2.1}$$

Traditional language modeling systems are mainly symbolic, which means they commonly rely on manually crafted rules [82]. More recent studies have shifted the research direction towards statistical techniques, such as n-grams, Hidden Markov Model (HMM), and neural architectures, as the preferred approach for building language models [36]. As our work primarily leverages neural language models, the subsequent sections will introduce the technical concepts underlying neural language modeling.

The process of text generation using neural language models can be described in three fundamental steps. Firstly, the input word sequence is mapped into vector representations, also known as embeddings, on a token-by-token basis. These embeddings capture the semantic or contextual information of each token. Secondly, the language model processes these embeddings, extracting intricate patterns and dependencies to generate a more condensed and higher-level representation of the input sequence. Lastly, this refined representation is used to predict the probability distribution over the vocabulary, indicating the likelihood of each word being the next in the generated text. By sampling or selecting the word with the highest probability, the next token is generated.

2.3.1. Word Embeddings

A crucial consideration in developing neural network-based models is determining an effective approach for representing information. Using word embeddings has emerged as a practical and successful strategy for neural language models. Word embeddings are distributed representations of words in a continuous vector space. These vectors are intended to capture semantic, syntactic, or contextual information embedded in the respective words. Several popular word embedding techniques are developed over the years.

Traditionally, embedding vectors are realized as **Static Embeddings**, vocabularies structured like dictionaries. In a static embedding, each token is assigned a distinct vector representation, either manually decided or learned. Embeddings built using this technique include, e.g., simple One-hot Encoding, Word2Vec, and GloVE Mikolov et al. [52] and Pennington et al. [58]. Static word embeddings can be directly used as input features for other models, providing ready-to-use information and improving model performance. Once trained, they are computationally efficient to use, with little or no fine-tuning required. However, a few limitations are yet to be improved upon.

When faced with new vocabulary, static embeddings typically represent all unknown words with an out-of-vocabulary (OOV) token. This approach stems from the fixed vocabulary size of static embeddings and can lead to performance limitations. Additionally, addressing polysemy or homonymy poses another challenge when employing static embeddings. Since the objective of word embeddings is to capture a word's meaning in a singular vector representation, multiple word senses will be compressed into the same embedding vector, leading to ambiguity and impacting performance outcomes.

To tackle the limitations of static word embeddings, **Language Model Embedding (LM Embedding)**, also called **Contextual Embeddings**, are proposed. Different from static embeddings, LM Embeddings are produced by a neural language model on each input sequence. This allows the vector representation of a specific token to vary based on its current context. A well-known example of LM Embeddings is the **Embeddings from Language Models (ELMo)**. ELMo is one of the first approaches to produce contextual word embedding using a neural language model [59]. It utilizes an architecture based on character-level convolutions and bi-directional LSTMs. Using character-level convolutions also allowed ELMo to handle OOV words more effectively.

2.3.2. Encoder-Decoder Architecture

An Encoder-Decoder architecture is widely adopted for constructing language models. The Encoder network is responsible for processing the embeddings, extracting semantic or contextual information, and capturing a higher-level representation of the input sequence. On the other hand, the Decoder network utilizes the condensed representation obtained from the Encoder to generate a probability distribution over the vocabulary, thereby predicting the subsequent token following the input sequence.

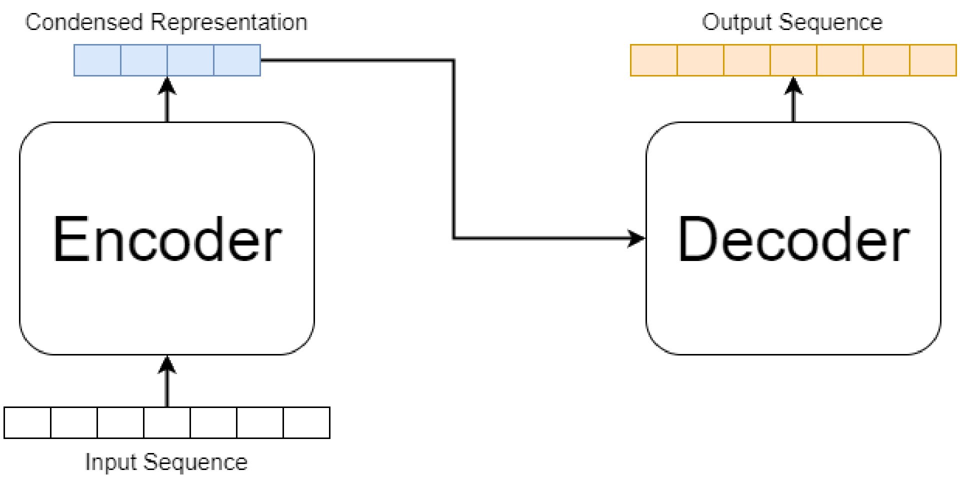

Figure 2.1.: A much-simplified illustration of an Encoder-Decoder architecture.

To achieve this, the Encoder-Decoder model employs a series of iterative steps. After sampling or selecting the next token, the model incorporates it into an extended input sequence. This extended input sequence serves as the new input for the model, allowing it to make predictions for the next token in a step-by-step manner. This process continues iteratively until a stopping criterion is met. Common stopping criteria include reaching a maximum sequence length or generating a designated stop token, indicating the completion of the generation process.

The versatility of the Encoder-Decoder architecture lies in its adaptability to various neural network implementations. Different types of neural networks can be employed as the Encoder or the Decoder, each offering unique advantages and capabilities. Researchers have explored and utilized diverse architectures, such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and, more recently, the Transformer model, which has gained significant attention and achieved remarkable results in natural language processing tasks. The choice of neural network for the Encoder and Decoder components depends on the specific requirements of the task at hand, as well as the desired trade-offs between computational efficiency, modeling capacity, and the ability to capture long-range dependencies within the data. A simplified illustration of the Encoder-Decoder architecture is shown in figure 2.1.

2.3.3. Transformer

Traditional RNN-based models are known to suffer from limited memory capacity and slow computation due to their architecture. RNNs process input tokens sequentially, where the hidden states are dependent on previous computations. This makes it challenging to implement parallel computing. Utilizing the idea of attention mechanism, first proposed by Bahdanau et al. [10], Vaswani et al. [74] introduced the Transformer architecture, replacing the recurrent network with so-called self-attention layers.

The architecture of the Transformer is shown in Figure 2.2. As illustrated, the transformer architecture employs an Encoder-Decoder structure that consists of two special ingredients: (1) the **Multi-head Attention** Layer and (2) the **Positional Encoding**. The encoder and decoder are each a stack of multiple identical building blocks comprised of sub-layers.

Each layer in the encoder contains two sub-layers: a multi-head self-attention mechanism and a simple position-wise fully connected feed-forward network. There is a residual connection around each of the two sub-layers, followed by layer normalization. In addition to the two sub-layers in the encoder, the decoder inserts a third sub-layer, the masked multi-head attention. The masked multi-head attention layer ensures that at inference time, the decoder does not include outputs from the future that are not yet generated. Similar to the encoder, residual connections and layer normalization are implemented.

Self-Attention

The attention mechanism is introduced by Bahdanau et al. [10] in the context of neural machine translation for improving performance regarding the long-distance dependency issue of traditional Seq2Seq models. As RNN-based models process input tokens sequentially and compress information of the entire sequence into a fixed-length representation, information loss is unavoidable.

The attention mechanism helps Seq2Seq models encode long sequences efficiently by allowing the decoder to "attend" to different positions of the input sequence, avoiding

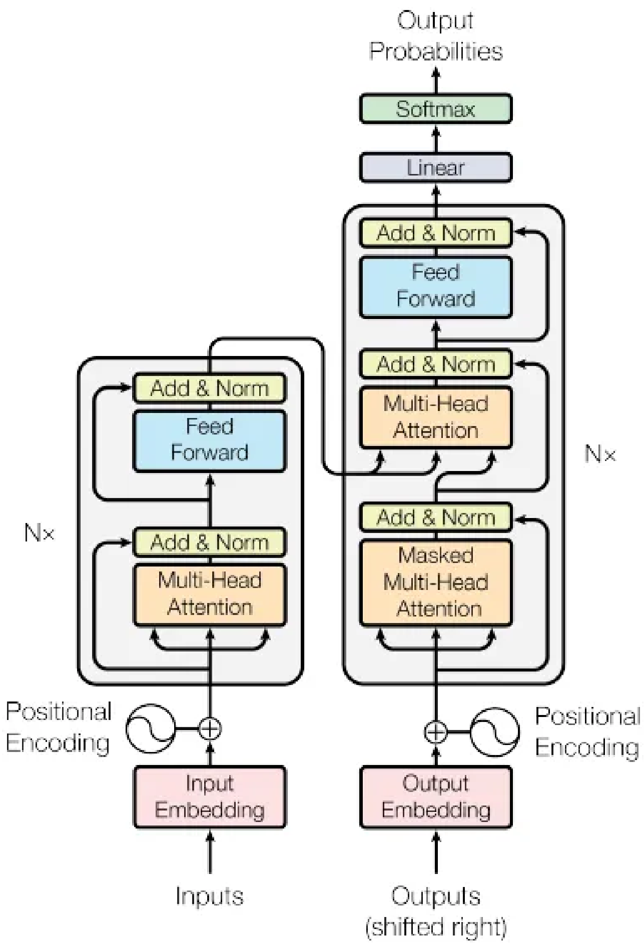

Figure 2.2.: The original illustration of the Transformer model architecture [74].

the need to compress the entire input sequence. Instead, the encoder produces a hidden representation h_i for each token and concatenates them into a hidden state $H = [h_1, ..., h_n]$, where n is the length of the input sequence. During inference time, the decoder computes an attention score for each encoder hidden representation h_i using a pre-defined score function $score(s_t, h_i)$, where s_t is the decoder hidden state at timestep t. The attention scores are then normalized into attention weights using a softmax function:

$$a_{t,i} = \frac{\exp(score(s_t, h_i))}{\sum_j^n \exp(score(s_t, h_j))} \tag{2.2}$$

Using the calculated weights, the decoder computes a context representation c as the weighted sum of the encoder hidden representation h_i:

$$c_t = \sum_i^n a_{t,i} * h_i \tag{2.3}$$

The process of computing the weighted sum using the attention mechanism is also called "attention pooling". This context representation c_t can be seen as a summary of the input sequence for the time step t and encodes which parts of the input sequence are more relevant for the output token that is currently being generated.

Taking the concept of encoding relevance by "attending" to different parts of a sequence, Vaswani et al. [74] introduced self-attention. Different from the original attention mechanism, where the decoder attends to encoded hidden states from the encoder, the self-attention allows each token to "attend" to all other tokens and therefore reflect the relations between different tokens in the same sequence. Hence the name "self-attention": a sequence attends to itself.

In Transformer, an input sequence is first transformed into a sequence of token embeddings $X \in \mathbb{R}^{n \times d} : X = [x_1, ..., x_n]^T$, where n denotes the total number of tokens and d the dimension of each token embedding. The sequence is further transformed into a Query matrix $Q = [q_1, ..., q_n]^T$, a Key matrix $K = [k_1, ..., k_n]^T$, and a Value matrix $V = [v_1, ..., v_n]^T$. The transformation is performed by multiplying the token embeddings by three learned weight matrices:

$$Q = XW^Q \tag{2.4}$$

$$K = XW^K \tag{2.5}$$

$$V = XW^V \tag{2.6}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times d_k}$ denote the weight matrices for the query, key, and value, respectively. The shapes of the three matrices are $Q, K \in \mathbb{R}^{n \times d_k}$, and $V \in \mathbb{R}^{n \times d_v}$, where d_k, d_V are the dimensions of the query, key, and value vectors. To encode the relation between tokens, an attention score S is computed for each token using dot product as the score function:

$$S = QK^T = \begin{bmatrix} q_1^T \\ q_2^T \\ ... \\ q_n^T \end{bmatrix} [k_1, k_2, ..., k_n] \tag{2.7}$$

$$= \begin{pmatrix} q_1k_1 & q_1k_2 & ... & q_1k_n \\ q_2k_1 & q_2k_2 & ... & q_2k_n \\ ... & & & \\ q_nk_1 & q_nk_2 & ... & q_nk_n \end{pmatrix} \tag{2.8}$$

where q_i, k_i are query and key vectors of the i-th token. This matrix $S \in \mathbb{R}^{n \times n}$ can be seen as a similarity matrix that measures the compatibility of each token with every other token in the sequence. Intuitively, the magnitude of the dot product between the query and key vectors tends to grow with their dimension k. Based on empirical results, Vaswani et al. [74] proposed to normalize the attention score based on the dimensionality d_k:

$$S_{normalized} = \frac{S}{\sqrt{d_k}} \tag{2.9}$$

Similar to the original attention mechanism, the scaled dot-product attention scores are converted into attention weights by applying a softmax function:

$$A = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{2.10}$$

The final output of the self-attention mechanism is a weighted sum of all value vectors, using the attention weights as weighting factors. Formally, the attention operation is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.11}$$

Multi-Head Self-Attention.

When considering a set of query, key, and value vectors capturing specific relations among tokens, drawing an analogy to convolution filters employed in image processing seems intuitive. In CNNs, individual convolutional filters are dedicated to capturing one type of feature. Therefore, CNNs employ multiple filters simultaneously, aiming to capture a set of different feature representations. In a similar manner, utilizing multiple sets of query, key, and value vectors to encode various types of relations can be beneficial. Each of these sets is referred to as a "head". Hence, the term "multi-head self-attention" is self-explanatory.

The multi-head self-attention consists of h different self-attention layers in parallel. Each attention head H can learn independent linear transformations for its query, key, and value vectors through attention pooling. The resulting outputs H_i from each attention head are computed using a different set of weight matrices W_i^Q, W_i^K, W_i^V, which are subsequently concatenated. To obtain the final output, the model applies a final linear transformation to the concatenated output vectors. Formally, multi-head self-attention can be defined as:

$$Multihead(Q, K, V) = Concat(H_1, H_2, ..., H_i)W^O \qquad i = 1, ..., h \tag{2.12}$$

$$H_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2.13}$$

By employing a set of heads simultaneously, multi-head self-attention provides advantages in both the efficiency and expressiveness of the model. Firstly, attention pooling can be parallelized highly efficiently, resulting in comparable costs for computing h attention heads and single-head attention. Secondly, using multiple attention heads allows the model to map different relationships between tokens onto multiple representation sub-spaces, using different sets of query, key, and value weight metrics. This increases both the expressiveness and the attention span to a wider range of sequences. This improves model performances, e.g., when dealing with sequences where both long-distance and short-distance dependencies are important [21].

Positional Encoding

The sequential nature of textual data is an important aspect of NLP and has been the motivation for RNN architecture. As the attention mechanism process each token independently from the other tokens, this temporal aspect of language is lost. To compensate for this lack, Vaswani et al. [74] introduced Positional Encoding, which provides the model with information about the relative positions of each token in the input sequence.

The positional encoding is added to the input embeddings for each token before they are fed into the first layer of both the encoder and decoder. This results in different token embeddings for the same token depending on its relative position in the sequence. The positional encodings in the original Transformer paper are calculated by applying sine and cosine functions:

$$PE(pos, 2i) = sin(\frac{pos}{10000^{2i/d}}) \tag{2.14}$$

$$PE(pos, 2i+1) = cos(\frac{pos}{10000^{2i/d}}) \tag{2.15}$$

where pos is the position of the token in the input sequence, i is the dimension in the embedding vector, and d is the dimension of the model. Figure 2.3 shows a visualization of the positional encoding.

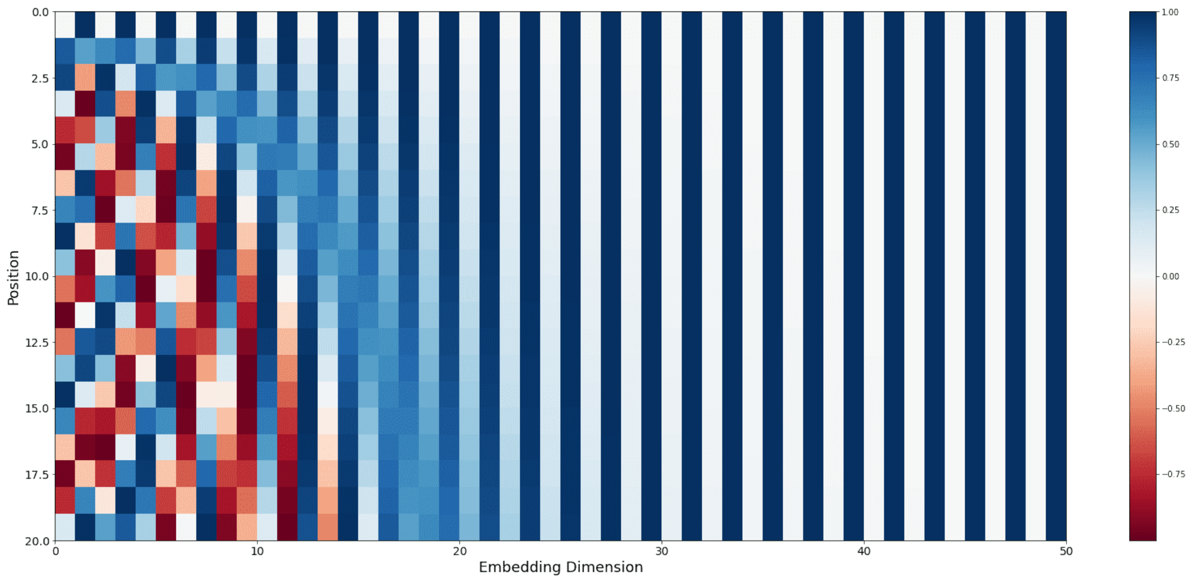

Figure 2.3.: A visualization of positional encoding for an input sequence of 20 tokens length and 50-dimensional embeddings, generated using a TensorFlow tutorial NoteBook [72].

2.3.4. Pre-trained Language Models

Language model pre-training has become increasingly promising in NLP [73]. With the introduction of the transformer model [74], building and training deep neural language models became significantly more cost-efficient [53]. By pre-training on a self-supervised task, e.g., language modeling, transformer-based models can leverage the large amounts of unlabeled corpora available to learn a generic representation of language, which can be easily transferred to other NLP tasks via transfer learning [85]. The common setup to pre-train language models is to first pre-train them on large unlabeled text data using a self-supervised training objective and then fine-tune the models on task-specific labeled data for transferring the knowledge learned to downstream tasks [21].

In contrast to traditional models that utilize existing word embeddings, pre-trained language models obtain the contextual language model embeddings during the pre-training and therefore require less manual feature engineering [59]. By exploiting the advantages of transfer learning, pre-trained models require a relatively small amount of annotated data while demonstrating better generalization ability [38].

As the models, as well as the datasets for pre-training, grow increasingly larger in scale, pre-training requires high computation power and sophisticated model architectures. However, pre-trained models can be applied easily to specific tasks with little to no fine-tuning. Utilizing already pre-trained models available on various NLP platforms and fine-tuning them for specific applications is comparatively cost-efficient, hence, making it the standard approach for NLP research [63].

Below we overview the pre-trained models that are used in this thesis, i.e., BERT, GPT, and BART.

BERT

With the success of the Transformer model, Devlin et al. [26] proposed a transformer-based architecture called **Bidirectional Encoder Representations from Transformers (BERT)**. As the name suggests, BERT is designed for creating bidirectional language model embeddings. BERT employs an encoder-only architecture based on the transformer encoder and is pre-trained on the BooksCorpus [90] and the English Wikipedia.

The BERT model is considered bidirectional due to its novel training objective, the Masked Language Model (MLM) objective. In the MLM training objective, a portion of the input tokens are randomly masked, and the model is tasked with predicting the original tokens based on the surrounding context. This training approach enables BERT to learn contextualized representations of words that take into account both preceding and succeeding contexts. In addition, BERT is simultaneously pre-trained on the Next Sentence Prediction (NSP) objective. In the NSP training objective, the model is provided with pairs of sentences as input and trained to discern whether the second sentence naturally follows the first. This objective aims to enhance BERT's comprehension of sentence-level context.

BERT achieved state-of-the-art results across various NLP tasks, with significant improvements to the previous high scores. Besides the success, Devlin et al. [26] demonstrated that BERT is particularly effective in transfer learning. Once the model is pre-trained, it can be fine-tuned on a specific task with a small amount of task-specific training data to achieve state-of-the-art results. Currently, BERT is commonly used for NLP tasks involving language understanding with complete context, e.g., sentiment classification, named entity recognition, and question answering [21].

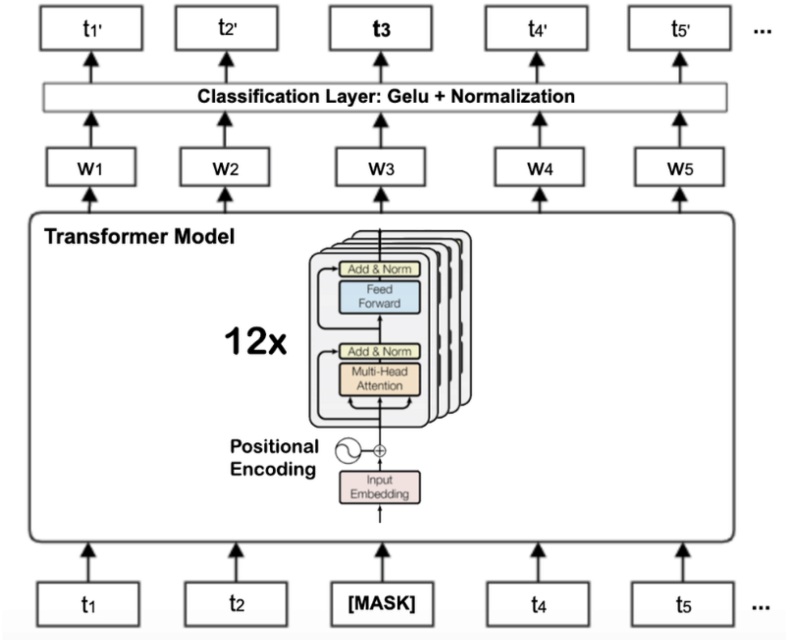

Figure 2.4.: A visualization of BERT-base model architecture with 12 encoder blocks and 12 self-attention-heads [40].

GPT

To explore the capability of transformer architecture, Radford et al. [61] introduced the **Generative Pretrained Transformer (GPT)**, a language model designed to learn a universal representation, generate coherent and contextually relevant sentences, and be fine-tuned for a wide range of tasks with little adaptation.

To leverage the large unlabeled text data available, GPT is designed as an autoregressive language model, which means that it generates an output token based on the entire input sequence at each time step. This allowed GPT to be pre-trained in an unsupervised fashion, with the objective of predicting the next word in the sequence, as each sequence is its own label. The autoregressive setting is similar to the task performed by the transformer decoder. Thus, GPT employs a decoder-only architecture with a stack of transformer decoder blocks. The architecture of the original GPT model is illustrated in Figure 2.5.

The general task-agnostic GPT model outperformed various existing language models crafted and trained specifically for each task, significantly improving upon the state-of-the-art performance. Since the publication and success of the first GPT model [61], several following iterations of GPT were published. Each generation was expanded on the basis of the previous version, with deeper and larger layers, more training data, and more sophisticated learning methods. E.g., GPT-2 [62] incorporated multitask and meta-learning settings, and the recent GPT-3.5 and GPT-4 employed RL to include human feedback in training [56].

As we focus on the pre-trained models used in this thesis here, we introduce the GPT-2

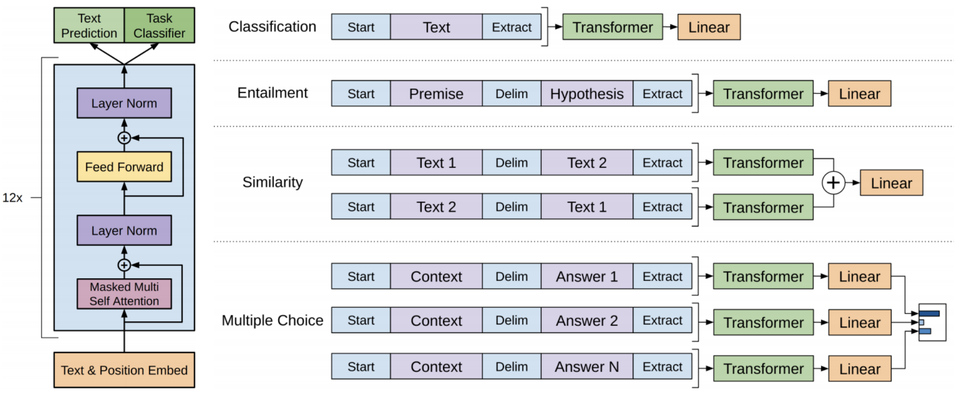

Figure 2.5.: **(left)** Transformer decoder architecture and training objectives used for GPT-1 with 12 decoder blocks and 12 attention heads. **(right)** Input transformations for fine-tuning on different tasks [61].

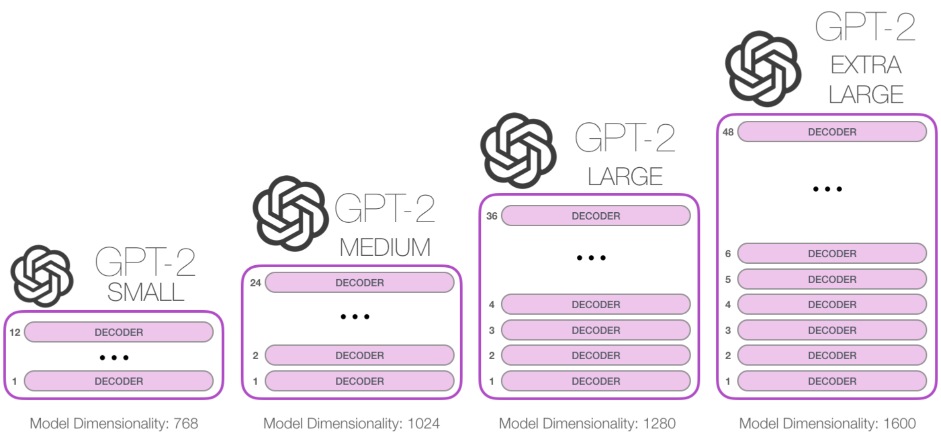

Figure 2.6.: An overview of available GPT-2 model sizes provided by Alammar [2].

model in greater detail. As an expansion to the original GPT-1 model, GPT-2 is available in four sizes, ranging from 12 to 48 stacks of transformer decoder blocks. To learn more general purposed language models with pre-training, GPT-2 employed the multitask learning setting. Instead of training a new model for each task, GPT-2 is designed to learn a conditioned probability of output on the desired task. Under this setting, the language modeling loss is formulated as:

$$L_1(W) = \sum_i \log P(w_i | w_{i-k}, ..., w_{i-1}, T; \Theta) \tag{2.16}$$

where w_i denotes a token in a dataset W, k the size of context window, T the task conditioning, and Θ the learnable parameter. Task conditioning can be e.g. in the form of prompting, where the task is specified by natural language instructions in the input sequence [21]. An example of the machine translation task can be structured as *"Translate from German to English,*

<German Sentence>, <English Sentence>".

The expansion in both the model and training data dimensions, coupled with the implementation of a multitask training setting, has propelled GPT-2 towards a generic language system. GPT-2 models demonstrated remarkable zero-shot capability in a wide range of tasks, outperforming various task-specific language models without fine-tuning [21].

BART

With the empirical success of masked language models, Lewis et al. [44] proposed a novel transformer-based masked language model that combines the strengths of both Bidirectional Auto-Encoding models (like BERT) and Auto-Regressive models (like GPT). Auto-regressive models predict each word in a sequence based on the previous words, which allows them to generate fluent, coherent text. However, their performance is limited on tasks that require understanding the full context of a sentence. Auto-encoding models, on the other hand, are trained to predict masked words in a sequence based on all the other words. This allows them to fully utilize the context of the complete input sequence. However, they are typically trained to predict only a small fraction of the words in the sequence, which can limit their ability to generate text.

The **Bidirectional and Auto-Regressive Transformers (BART)** was designed to combine the strengths of these two types of models by employing an encoder-decoder architecture, including both the transformer encoder and decoder blocks. Different from the original transformer model, BART employs a training setting akin to a denoising autoencoder. The original transformer is trained on sequence-to-sequence tasks, where the encoder extracts sentence representation, and the decoder reconstructs the target sentence. In contrast, the BART encoder is fed with corrupted input sequences while the decoder is trained to reconstruct the uncorrupted version. The input sequences are corrupted in two approaches [44]: (1) randomly shuffling the order of the original sentences and (2) replacing arbitrary length spans of text (including zero length) with a single mask token. The reconstruction process for a target sequence X can be formulated as:

$$L(X) = -\sum_i \log P(x_i | x'_1, ..., x'_n; \Theta) \tag{2.17}$$

Here, $P(x_i | x'_1, ..., x'_n; \Theta)$ denotes the conditional probability of the i-th token given the corrupted sequence $X' = [x'_1, ..., x'_n]$, with Θ being the learnable model parameters.

This novel training methodology enables BART to learn a more robust representation of natural languages. The denoising autoencoder architecture allows it to use the complete context from the input sequence like an auto-encoding model, while it also predicts each word based on the previous words, like an auto-regressive model. BART demonstrated outstanding performance on generative tasks that require a deep understanding of context, such as summarization, question answering, and translation [45].

2.4. Evaluation Metrics

Evaluation is challenging for dialogue systems, as the criteria for valid responses are diverse. In this section, we introduce commonly employed automatic metrics for evaluating dialogue systems, while also acknowledging their inherent limitations. Automatic metrics play a vital role in the domain of machine learning, providing objective, quantitative measures for evaluating and comparing model performances. Notably, they offer advantages over human evaluation, such as cost-efficiency and quick computation, enabling rapid and standardized assessments of a model's proficiency. We roughly divide automatic metrics into two categories: similarity-based metrics and classification-based metrics.

Similarity-based Metrics

The intuition behind similarity-based metrics is simple. Given a reference sentence, a generation model should produce outputs that are similar to the reference. Some common similarity-based metrics include, e.g., BLEU and ROUGE. BLEU [57] is primarily used for evaluating machine translation tasks, but can also be applied to language modeling even if it's criticized [49]. It measures the similarity between the generated output and the reference text using a modified n-gram precision. Higher BLEU scores indicate better performance in terms of text similarity. Similar to BLUE, ROUGE [46] measures the word overlap between the predicted output and the reference text using various metrics, such as ROUGE-N (n-gram overlap), ROUGE-L (longest common subsequence) and ROUGE-S (skip-bigram co-occurrence statistics). E.g., in the domain of counter-speech generation, Zhu and Bhat [89] implemented the BLEU score and the ROUGE score for measuring syntactic similarity between generated counter-speech and the reference responses. This similarity score is used as an indicator for assessing the **Relevance** of generated responses. Saha et al. [64] used BLEU-29 and METEOR [65] to measure the similarity of the generated counterspeech to the reference responses, and Tekiroglu et al. [71] implemented the BLEU score as well for measuring lexical generation performances.

The traditional similarity-based metrics relied on hand-crafted rules that measure only the surface-level similarity between sentences[67]. With the advances of large language models (LLM), novel metrics based on contextual embeddings from large pre-trained Language Model (LM)s have been explored [39]. The idea is to compute the similarity between the extracted latent representations instead of the surface-level features. E.g., metrics like RUBER [70] and BERT-RUBER [32] utilize language models to compute the relevancy of a response regarding a given query. Tekiroglu et al. [71] also included the BertScore [86] for assessing semantic generation performances of counter-speech generation models.

Classification-based Metrics

Most similarity-based metrics are developed with specific tasks in mind, such as text summarization or machine translation, where an apparent target reference is given as the label, and a significant overlap between valid generation output and reference is assumed. This is a strong assumption and is not always valid in other tasks, like dialogue generation [47]. Dialogue generation is similar to our setting, as counter-speech generation simulates a conversational environment where an agent gives a response to the initial message. In such scenarios, reasonable responses don't necessarily share words or semantic meanings.

In comparison to similarity-based metrics, text classification models can directly assess specific characteristics without relying on reference texts. Therefore, they are often recruited as scoring systems for language models. E.g., Gao et al. [29] employs an ensemble of multiple GPT-2 based models to predict whether the response is engaging and human-like. Saha et al. [64] utilized a classifier of linguistic acceptability to evaluate the grammaticality of generated outputs and fine-tuned various transformer-based classifiers for detecting a set of attributes (toxicity, emotions, and politeness) of the outputs. By leveraging classification models, researchers were able to quantify the extent to which the generated responses successfully captured these desirable qualities. This direct assessment allows for targeted evaluations and facilitates the identification of areas for improvement in language generation tasks.

3. Related Works

In this chapter, we focus on exploring three areas that are relevant to the counter-speech generation. Firstly, we introduce publically available hate- and counter-speech datasets in Section 3.1, which can be effectively employed for training and evaluating automated counter-speech generation pipelines. Subsequently, we present an overview of existing approaches tackling automatic counter-speech generation in Section 3.2. Lastly, Section 3.3 examines a relevant work in the field of response generation, specifically focusing on generating contextually appropriate responses for app reviews. This work serves as the inspiration for our proposed framework.

3.1. Hate- and Counter-Speech Datasets

To develop and evaluate language models for generating counter-speech and classifiers for evaluating the generation results, we utilize several datasets containing either labeled instances of hate speech or hate- and counter-speech pairs. In the following sections, we introduce the structure of each selected dataset briefly.

3.1.1. A Benchmark Dataset for Learning to Intervene in Online Hate Speech

Conversation	Hate Speech	Human-Written Intervention Responses
1. User 1: United Kingdom: 'Schoolboy, 15, given detention for backing UKIP during classroom debate' 2. User 2: The education system is full of re***ds! Yes, most school teachers are ret***ed lefties! Teach your children to laugh at these ret***ed lefties! 3. User 3: Asking a teacher to not be a leftist is like asking a medieval munk to question the Pope. 4. User 4: The Jews are like Sjws, they infest everything.	2, 4	➢ Use of this language is not tolerated and it is uncalled for. ➢ Use of the slurs and insults here is unacceptable in our discourse as it demeans and insults and alienates others. ➢ I recommend that you research the holocaust, you might change your opinion.

Figure 3.1.: An example of the aggregated data. The first column shows texts from the conversation with indexes added to each post. Indentations indicate the reply structure. The second column shows indexes of the human-labeled hateful post. Each bullet point in the third column is a human-written response [60].

Qian et al. [60] introduced two benchmark datasets containing fully labeled conversational segments with human-written intervention responses.[1] The two datasets are collected from

[1]https://github.com/jing-qian/A-Benchmark-Dataset-for-Learning-to-Intervene-in-Online-Hate-Speech

Gab and Reddit, while the labels and intervention responses are manually created by Mechanical Turk workers via crowdsourcing. The Gab dataset consists of 11825 conversation segments and the Reddit dataset 5020. Figure 3.1 shows an example conversation segment.

The benchmark datasets from Qian have proven to be useful for training generative models that generate counter-speech. In the own work by Qian et al. [60], the author experimented with **Seq2Seq**, **VAE**, and **RL** to generate hate speech intervention. In addition, further studies on hate speech classification and intervention have been conducted utilizing these datasets [89] [64].

3.1.2. Counter-Narratives Datasets to Fight Hate Speech

The Counter-Narratives Datasets consists of four curated, well-documented datasets, created between 2019 and 2022.[2] Each of the four datasets provides high-quality data focusing on different aspects of online hate- and counter-speech conversations, enabling developments in various novel counter-narrative generation methods.

COunter NArratives through Nichesourcing

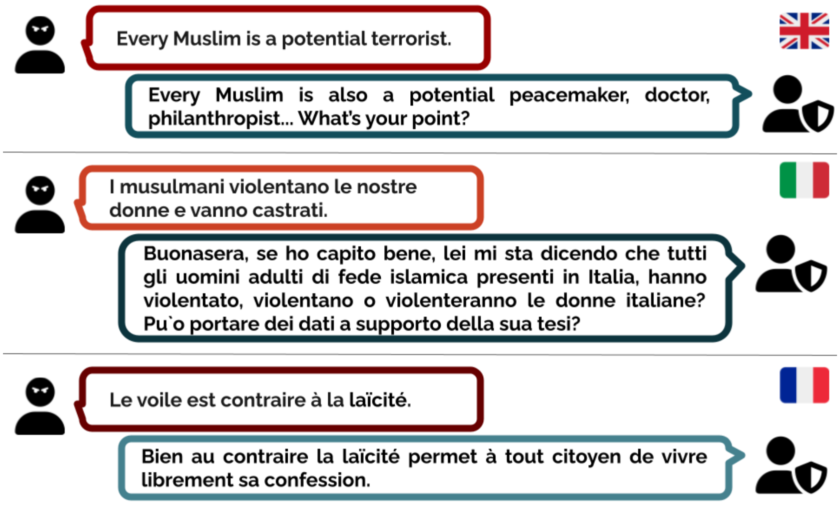

Figure 3.2.: An example of the hate- and counter-speech pairs from the CONAN dataset. One of the three pairs is the originally collected source data, and the other two are augmented through translation [23].

The first dataset was introduced by Chung et al. [23], called **COunter NArratives through Nichesourcing (CONAN)**. The CONAN dataset contains 4078 multilingual and expert-based hate- and counter-speech pairs in English, French, or Italian, with the topic of hate being Islamophobia. Each sample is annotated with expert demographics, hate speech sub-topic, and counter-speech type. By utilizing translation from French/Italian to English and

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

paraphrasing, the total number of hate- and counter-speech pairs is 14988. Figure 3.2 shows an example of a hate- and counter-speech pair in three languages.

In contrast to many publicly available NLP datasets, the CONAN dataset is completely hand-crafted by experts. Chung et al. [23] proposed a data-creation pipeline where trainers and operators from non-governmental organizations (NGOs) are recruited. These experts were given specific instructions to either write prototypical Islamophobic hate texts or to construct counter-speech in response to existing hate messages. NGO operators are professionally trained to tackle hated online, following guidelines to create fact-bounded information using non-offensive language. By utilizing human expertise, this approach aims to ensure the dataset's quality, reliability, and credibility.

Multi-target CONAN

The Multi-target CONAN dataset was the first addition to the original CONAN dataset, created by Fanton et al. [27]. Compared to the original CONAN dataset, the Multi-target CONAN dataset emphasizes on the variety of targeted demographics. It contains 5003 hate- and counter-speech pairs in English, covering 8 target demographics, including *DISABLED*, *JEWS*, *LGBT+*, *MIGRANTS*, *MUSLIMS*, *PEOPLE OF COLOR (POC)*, *WOMAN* and *OTHER*.

For constructing this dataset, Fanton et al. [27] proposed a human-in-the-loop data collection methodology. The proposed framework is shown in Figure 3.3. A GPT-2 language model was selected as the author module, first fine-tuned on a seed dataset of hate- and counter-speech pairs. The author module generates new counter-speech response candidates while human reviewers filter and post-edit them. The reviewed samples are then added back to the training data to further fine-tune the author module, improving the generation quality. This approach aims to increase the data collection efficiency while producing high-quality, human-approved data. Figure 3.4 shows an example hate- and counter-speech pair.

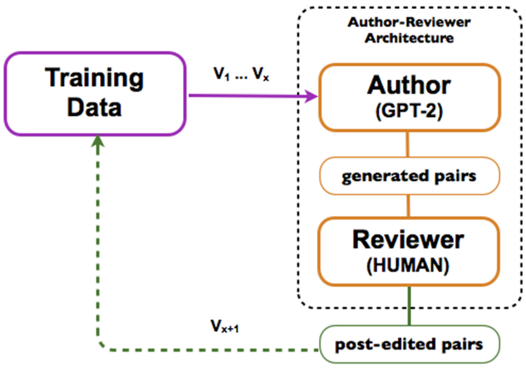

Figure 3.3.: The author-reviewer in the loop configuration. The author module produces HS/CN candidates, and the reviewer(s) validate and eventually post-edits them. New examples are added to training data at each loop, and the author is fine-tuned from scratch [27].

Figure 3.4.: A example hate- and counter-speech pair in Multi-target-CONAN dataset. The text in red represents posts from a hater and the text in blue is the expert-reviewed counter-narratives [27].

DIALOCONAN

As the previous two datasets provide hate- and counter-speech pairs in a single-turn fashion, Bonaldi et al. [13] introduced the **DIALOCONAN**, a multi-turn dialogues dataset.

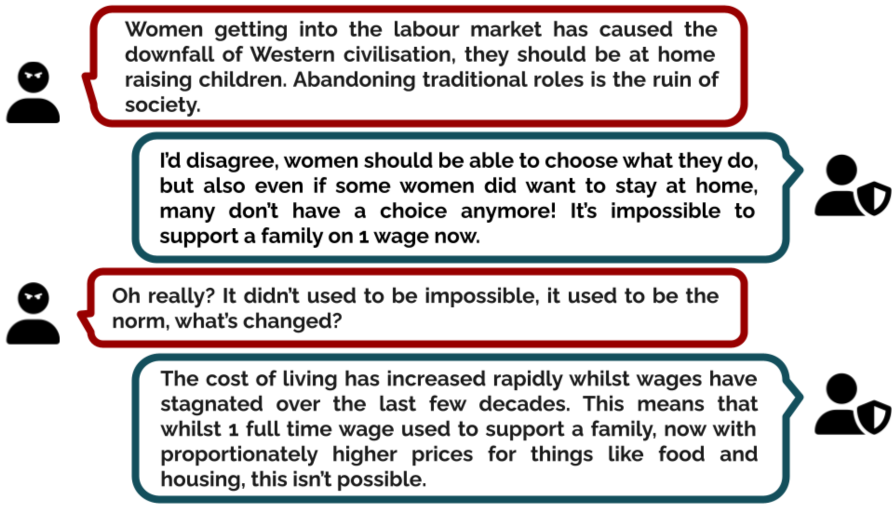

Figure 3.5.: A 4-turn example dialogue from the DIALOCONAN dataset. The text in red represents posts from a hater, and the texts in blue are counter-narratives from an NGO operator [13].

The **DIALOCONAN** dataset comprises 3059 conversation segments, either 4, 6, or 8 turns. Each dialogue represents a conversation between an online hater and an NGO operator. These dialogues cover 7 main target demographics (*JEWS*, *LGBT+*, *MIGRANTS*, *MUSLIMS*, *POC*, *WOMAN*, and *OTHER*). Figure 3.5 shows a 4-turn dialogue example. The samples are created via human expert intervention over machine-generated dialogues, an *author-reviewer* framework similar to the strategy used in [27].

Knowledge-grounded Hate Countering Dataset

The **Knowledge-Grounded hate countering dataset** [22] focuses on incorporating factual knowledge into counter-speech generation. It contains 195 hate- and counter-speech pairs, each coupled with the background knowledge used for constructing the counter-speech.

Chung et al. [22] employed a knowledge retrieval pipeline that automatically retrieves the relevant knowledge. The hate-speech/knowledge pairs are then provided to experts for constructing a suitable counter-speech, while the experts were instructed to utilize the content of the knowledge as much as possible. Figure 3.6 shows an example.

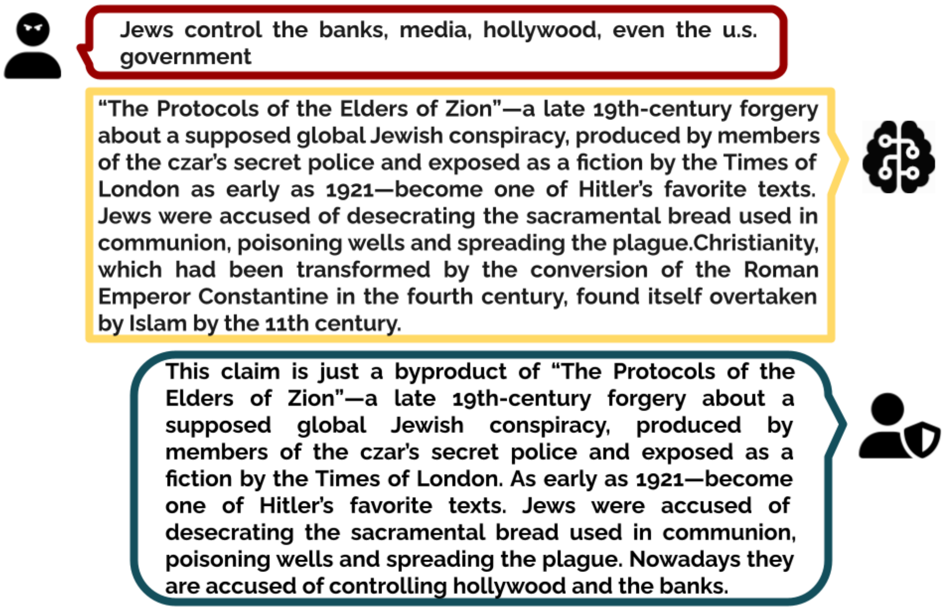

Figure 3.6.: An example of hate/counter-speech + background knowledge triplet in Knowledge-Grounded hate countering dataset. The text in red represents post from a hater, the text in blue is the corresponding counter narrative written by an expert and the text in yellow is the background knowledge for constructing the counter-speech [22].

3.1.3. EDOS: Explainable Detection of Online Sexism

The EDOS dataset[3] [41] is provided for a challenge on explainable detection of online sexism, organized by Rewire. The dataset consists of 20,000 entries, from which 10,000 are sampled from Gab and 10,000 from Reddit. Each entry is labeled by trained annotators or experts according to the scheme shown in figure 3.7.

[3] `https://github.com/rewire-online/edos`

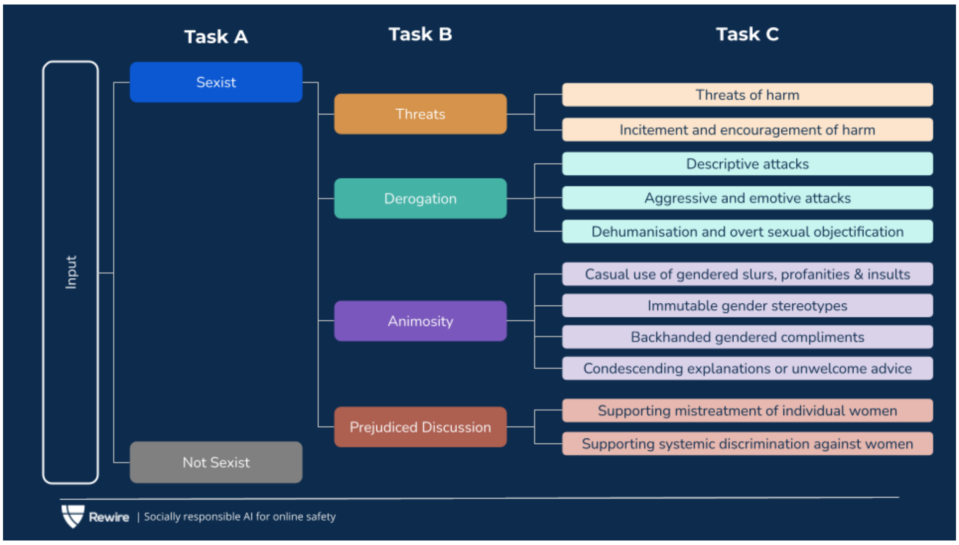

Figure 3.7.: The categorization scheme provided by EDOS challenge. Each entry is labeled as sexist or not sexist, the sexist entries are further categorized into four main categories and eleven sub-categories[41].

3.1.4. Summary

In this section, we briefly summarize the above-mentioned datasets in Table 3.1.

Name	Sample Size	Format	Hate Speech Source	Counter Speech Source	Language	Target
Qian et al. [60]	16845	Conversation Segments	Gab/Reddit	Crowd-Sourced	EN	generic
CONAN [23]	14988	Single-turn HS-CS-pair	Nichesourced Data Augmentation	Nichesourced Data Augmentation	EN/IT/FR	Muslims
Multi-target CONAN [27]	5003	Single-turn HS-CS-pair	Nichesourced	Computer-generated Human post edited	EN	Jews, LGBT+, Migrants, Muslims, POC, Woman, Disabled, Other
DIALOCONAN [13]	3059	Multi-turn HS-CS-pair	Nichesourced	Computer-generated Human post edited	EN	Jews, LGBT+, Migrants, Muslims, POC, Woman
Knowledge-grounded Hate Countering Dataset [22]	195	Single-turn HS-CS-Fact-pair	Nichesourced	Nichesourced	EN	Muslims
EDOS [41]	20000	Hate Speech	Gab/Reddit	-	EN	Woman

Table 3.1.: Summary of Hate- and Counter-Speech datasets.

3.2. Counter-Speech Generation

In the past years, a few studies have been conducted, aiming to generate high-quality counter-speech utilizing language models.

Qian et al. [60] evaluated the commonly used generation methods with their benchmark dataset, including a **Seq2Seq** encoder-decoder model proposed by Sutskever et al. [69], a **VAE** based on Seq2Seq, and a Seq2Seq model trained using **RL**.

Tekiroglu et al. [71] proposed an author-reviewer framework for generating counter-speech, where a pre-trained **GPT-2** model is fine-tuned to generate counter-speech candidates and human experts are recruited to post-facto edit the final responses.

Ashida and Komachi [4] investigated the potential of utilizing **prompting**. Finetuning large-scale pre-trained language models like GPT-2-large or GPT-3 requires huge amounts of high-quality data and computational power. To avoid finetuning, Ashida and Komachi [4] proposed to use a *few-shot* prompt design, including multiple examples of hate- and counter-speech pairs in the input sequence, to generate counter-speech directly with pre-trained models. They experimented this approach with **GPT-2**, **GPT-Neo** and **GPT-3**, using the biggest parameter size models for each GPT.

Besides these works, we introduce in the following sections two approaches in greater detail, as they were used in our study as two of the baseline models.

3.2.1. Generate, Prune, Select (GPS)

Zhu and Bhat [89] proposed a three stages pipeline to generate diverse and relevant counter-speech. The proposed pipeline consists of three individual modules: (i) Candidate Generation, (ii) Candidate Pruning, and (iii) Response Selection. The pipeline is illustrated in Figure 3.8.

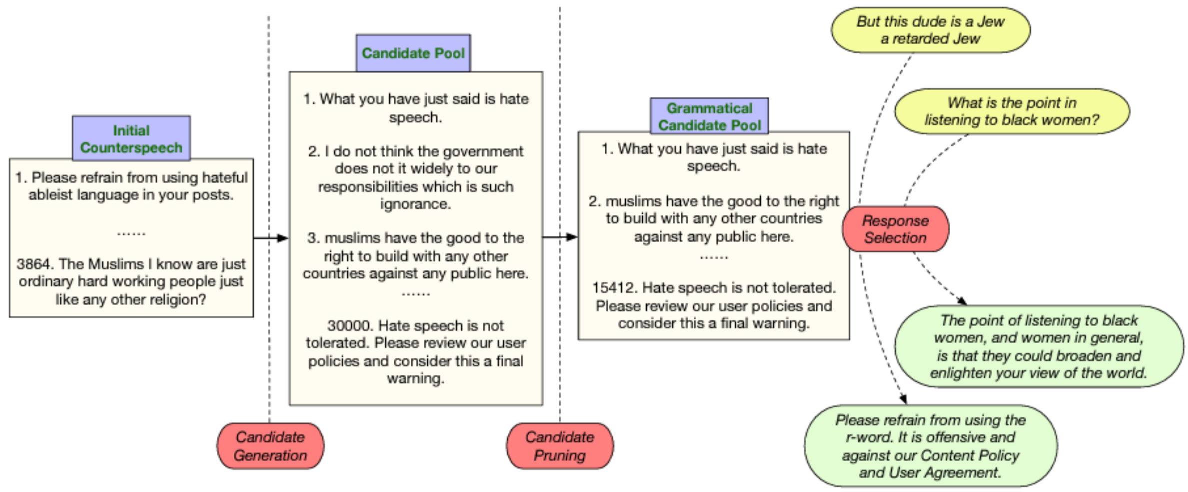

Figure 3.8.: Overview of the GPS pipeline. The red ovals correspond to the modules [89].

The core idea of this approach is to generate a large pool of grammatically correct counter-speech candidates and select the most fitting and relevant candidate for each input hate sequence.

Training Dataset

The GPS pipeline is trained on the combined dataset of the Benchmark datasets collected by Qian et al. [60] and the English portion of the CONAN dataset [23]. This combined dataset is then randomly split into 70-15-15 Train-Validation-Test sets.

Candidate Generation

The first step of the GPS pipeline is to create a counter-speech pool with diverse candidates. Using existing counter-speech instances as a starting point, a **RNN-based VAE** is trained to generate further counter-speech and extend the candidate pool. For this task, the VAE generator is trained on all available counter-speech instances $Y = [y_1, y_2, ..., y_n]$. The encoder and decoder have two layers with 512 nodes each and two highway network layers [68] for achieving more robust training behavior.

The training aims to maximize the lower bound of the likelihood $\mathcal{L}$ of generating the training data Y,

$$\mathcal{L} = -KL(q_\theta(z|y)||p(z)) + \mathbb{E}_{q_\theta(z|y)}[\log p_\theta(y|z)] \tag{3.1}$$

where θ represents the parameters of the generative model, while z represents a latent variable that follows a Gaussian distribution with a diagonal covariance matrix. The prior distribution is denoted by p and the posterior by q, and KL denotes the KL-divergence [43]. During inference time, the VAE model generates candidates by sampling noise from a standard Gaussian distribution and decoding the sampled noise (i.e., $\epsilon \sim N(0, 1)$).

Candidate Pruning

Since the generated candidates are not always grammatical, Zhu and Bhat [89] fine-tuned a pre-trained BERT model ('bert-base-cased') on the Corpus of Linguistic Acceptability (CoLA) [79]. This BERT model acts as the grammaticality classifier. All candidates are rated by this classifier, and the ungrammatical candidates are pruned.

Response Selection

With a large pool of grammatical, diverse counter-speech candidates, a response selector aims to pair each input hate sequence with the most relevant response. Zhu and Bhat [89] proposed a novel approach for selecting the final counter-speech.

Firstly, sentence embeddings are obtained for both the input hate sequence and the candidate counter-speech, utilizing pre-trained models like sentence transformers. Let e_x denote the embedding of the input hate sequence and e_y denote the embedding of the candidate counter-speech. A linear mapping $e'_y = (W + BI)e_y$ is learned using the hate- and counter-speech pairs in the training data so that the sum of the cosine similarities between e_x and e'_y is maximized. Using this transformation, the cosine similarity score between the input embeddings and the candidate embeddings can be considered a relevancy score for selecting the best candidate.

The GPS pipeline outperformed other baseline models, including a BART model fine-tuned on the same training dataset, in both automatic metrics and human evaluation, scoring high in Diversity, Relevance as well as Language Quality.

3.2.2. Counter-GEDI

Saha et al. [64] proposed a controllable approach for generating counter-speech with specific attributes. The idea behind this approach is to utilize **Generative Discriminator (GEDI)** to influence the next token probability during inference.

GEDI is presented by Krause et al. [42] as a decoding time algorithm to control the output from generation models. Assume we have a class-conditioned language model with a desired attribute c and an undesired attribute $\bar{c}$. E.g., "polite" is the positive label and the desired attribute, and "non-polite" is the negative and undesired attribute. The GEDI model is trained to learn the contrast between $P_\theta(x_1 : t|c)$ and $P_\theta(x_1 : t|\bar{c})$. This contrast is used to guide the generation model during inference, which gives $P_{LM}(x_1 : T)$. The GEDI models are integrated into the generation pipeline by implementing heuristic equations. Krause et al. [42] introduced a simple heuristic equation for controlling one attribute using a single GEDI model:

$$P_w(x_t|x_{<t}, c) \propto P_{LM}(x_t|x_{<t})P_\theta(c|x_t, x_{<t})^\omega \tag{3.2}$$

where ω is the control attribute. Saha et al. [64] extended the heuristic for using multiple GEDI models to control multiple attributes simultaneously:

$$P_w(x_t|x_{<t}, c_1, .., c_n) \propto P_{LM}(x_t|x_{<t}) \prod_{i=i,..,n} P_\theta(c_i|x_t, x_{<t})^{\omega_i} \tag{3.3}$$

where ω_i is the $i-$th control attribute for guiding the generation toward class c_i, using a GEDI model trained on the $i-$th attribute.

CounterGEDI Pipeline

The CounterGEDI pipeline consists of three major components: (A) a counterspeech generation model, (B) a set of GEDI models, and (C) an output token selector.

Part A is a conventional autoregressive generative model that takes the input hate sequence and the generated tokens from previous time steps and outputs the next token probability distribution. Saha et al. [64] fine-tuned a **DialoGPTm** model as the generation model. **Part B** is a set of GEDI models that control different attributes. Each GEDI model takes in only the generated tokens from previous time steps and computes the next token probability distribution based on its control attribute. Saha et al. [64] proposed to activate the GEDI models after 10 generation steps, allowing the generation model to produce an initial counter-speech segment. **Part C** represents the output token selector, where the next token probability distribution from the generation model and the GEDI models are weighted, and a final token is selected as the next token based on the combined probability distribution. The weights for each GEDI model are manually adjusted based on the importance of each control attribute.

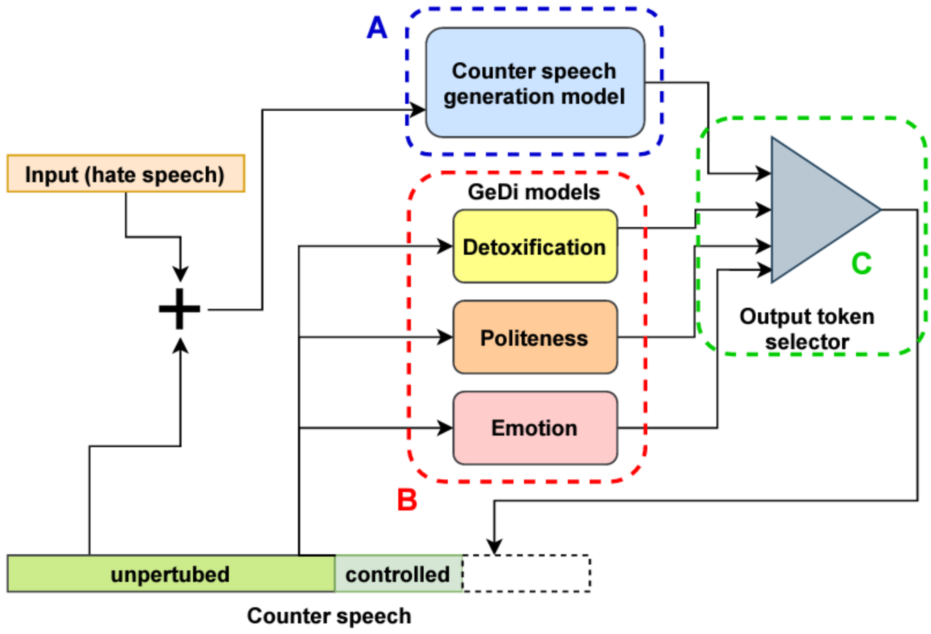

Figure 3.9.: An overview of the CounterGEDI pipeline setup. The counterspeech generation model and the GEDI models each produce a probability distribution for the next token, taking in either the concatenated sequence of hate speech and previously generated counter-speech or only the counter-speech. These distributions are weighted by the output token selector for choosing the next output token. The unperturbed part in the counterspeech is created without any control to provide the initial prompt to the GEDI model [64].

Training Datasets

For training the generation model, the Benchmark datasets from Qian et al. [60] and the English portion of CONAN [23] are selected. For each dataset (Gab, Reddit, CONAN), a separate generation model is fine-tuned using an 80-10-10 Train-Validation-Test Split.

Saha et al. [64] experimented with six control attributes, *politeness*, *toxicity*, *joy*, *fear*, *sadness* and *anger*. Each attribute is controlled by a separate GEDI model.

The *politeness* GEDI model is trained on the dataset published by Madaan et al. [48], where 1.39 million posts are labeled into nine politeness classes (P1-P9). Following the recommendation of the author, posts rated P9 are considered "polite", and other classes "non-polite".

The *toxicity* GEDI model is trained on a popular Kaggle dataset[4] containing text samples labeled as "toxic" or "non-toxic".

The other four GEDI models are trained on a large dataset [66] containing 416,809 posts

[4]Toxic Comment Classification Challenge: https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/rules

labeled with seven emotions ("sadness", "joy", "fear", "anger", "surprise", and "love"). Saha et al. [64] removed samples labeled as "surprise" or "love" as their numbers are too low for fine-tuning. While training for each emotion, the chosen emotion is labeled as the positive attribute, while all other emotions are labeled as negative, thus resulting in four binary GEDI models.

Saha et al. [64] showed that the CounterGEDI pipeline is capable of producing counter-speech with desired attributes while maintaining high language quality.

3.3. Response Generation

counter-speech generation can be considered a sub-task of dialogue/response generation. As this task is more intensively researched, we introduce the one publication from which we drew inspiration.

Zhang et al. [87] proposed a Transformer-based approach for automated app-review response generation called **TRRGen**. This approach incorporates the apps' categories and the review ratings as feature embeddings and concatenates them onto the input sequence embeddings, providing the model with additional information. Their results indicate that by incorporating these additional features, the model was able to generate higher-quality, more relevant responses and outperform state-of-the-art approaches to the task. Figure 3.10 shows the proposed architecture.

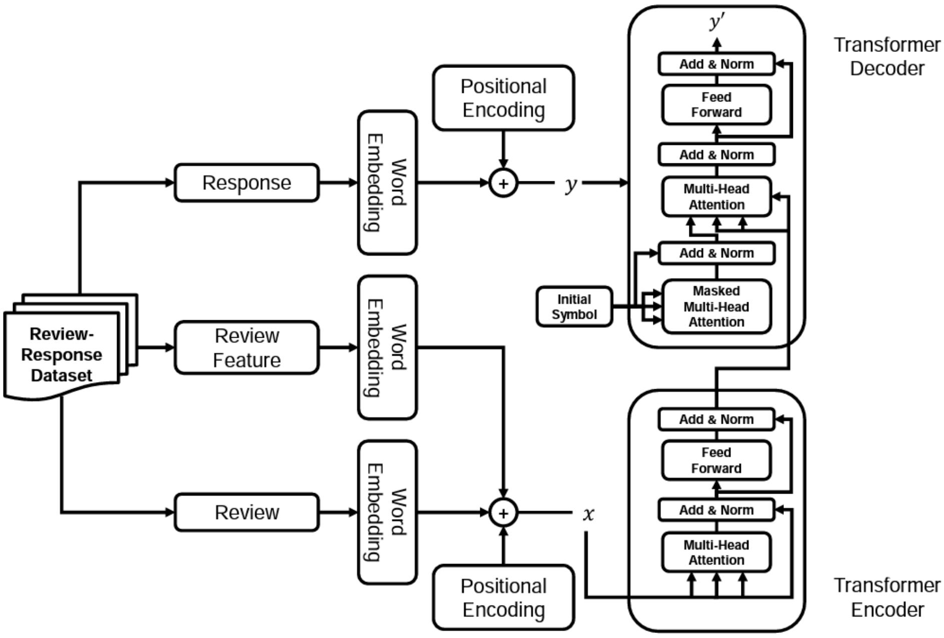

Figure 3.10.: An overview of the TRRGen architecture [87]. In addition to the conventional transformer architecture, Zhang et al. [87] incorporated extra information into the input, in this case, the app's category and the review's rating as **Review Features**.

The main idea here is to provide the model with additional information and thus help them generate more relevant responses. This approach is also valuable for language generation in other domains, especially in dialogue systems, where the relevancy of the response is crucial for the generation quality. In our work, we examine the effectiveness of this approach in the counter-speech generation scenario.

4. Methods

To answer the stated research questions, we designed experiments examining the effectiveness of currently available methods for counter-speech generation. In Section 4.1, we present our proposed framework for improving the relevancy of generated counter-speech responses regarding specific target demographics. Subsequently, Section 4.2 provides a detailed review of our data preparation pipeline. Section 4.3 introduces the procedure for our experiments, describing each step conducted for obtaining the results. Lastly, the evaluation pipeline is introduced in Section 4.4 and 4.5, covering the automatic metrics and human evaluation implemented in our experiments, respectively.

4.1. Proposed Framework

Besides testing existing techniques for generating counter-speech against sexist hate speech specifically, we also propose a transformer-based generative language model, aiming to improve the generation quality regarding the attribute **Relevancy**. Inspired by the work from Zhang et al. [87], we propose to fine-tune a transformer-based model that incorporates the target groups of the input hate speech as an additional feature.

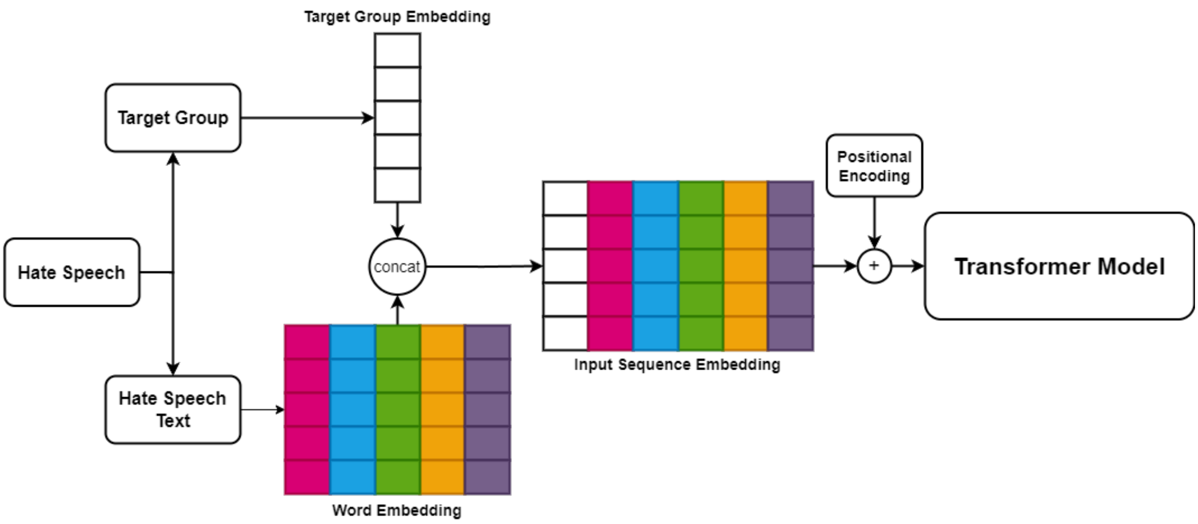

Figure 4.1.: Illustration of the proposed transformer-based counter-speech generation framework.

	Input Sequence
Original	[USER] Leg day is easy. Hot girls who wear miniskirts get asked out.
Autoregressive Model	<WOMEN> Hate Speech: [USER] Leg day is easy. Hot girls who wear miniskirts get asked out. Counter-Speech:
Seq2Seq Model	<WOMEN> [USER] Leg day is easy. Hot girls who wear miniskirts get asked out.

Table 4.1.: An example of the input sequences used for different models in combination with the target group embedding.

Target Group Embedding

To incorporate target group information of hate speech, we propose to adopt the 8 target groups from the Multi-target CONAN dataset as our defined classes. These target groups are ***Migrants***, ***People of Color (POC)***, ***LGBT+***, ***Muslims***, ***Women***, ***Jews***, ***Disabled***, and ***Other***.

For each individual target group, we define a new special token $<$ *Target-Group*> and add it to the model's vocabulary. This way, we can encode the target-group information into compact embedding vectors that can be further processed with input word embeddings. Similar to the proposed strategy in Zhang et al. [87], we concatenate the target group embedding at the beginning of each input sequence. This can be interpreted as adding the special target group tokens as the first tokens in the input sequence. An example of input sequences concatenated with target group token is shown in table 4.1, and the proposed framework is illustrated in figure 4.1.

4.2. Data Preparation

In this section, we present the data preparation performed for our experiments. From the publically available datasets, we selected a portion and constructed several custom datasets for the purpose of our experiments. The data selection process is described in Section 4.2.1. After the data selection, several preprocessing techniques are applied to the data, ensuring uniformity in the structure and quality of samples across various source corpora. The applied preprocessing for samples from each source corpora is provided in Section 4.2.2. Lastly, Section 4.2.3 provides an overview of all custom datasets we constructed.

4.2.1. Data Selection

For our experiments, we prepared two training datasets and three testing datasets for fine-tuning and evaluating selected models:

Custom Dataset is the main dataset we used for training or fine-tuning counter-speech generation models. This dataset is constructed using samples from the CONAN, Multitarget-CONAN, and DIALOCONAN datasets, containing 11,926 single-turn HS-CS pairs. The combined dataset is split randomly into train and test sets with 80% for training and 20% for testing. Finally, we end up with 9540 HS-CS pairs in the training set and 2386 HS-CS pairs in the testing set.

Topic Classification Dataset is a training dataset constructed for fine-tuning a topic classifier model, which is introduced later on in 4.4.5. The classifier focuses on classifying counter-speech into one or multiple of eight predefined target groups. Therefore, we included all counter-speech from the Multitarget-CONAN and DIALOCONAN datasets with their corresponding target labels. The final dataset contains 13314 counter-speech samples, randomly split into train and test sets with 90% for training and 10% for testing, making a 11982 samples training split and a 1332 samples testing split.

Small Custom Test Set is a balanced subset of the **Custom Dataset**. We selected 10 samples from each target group randomly, making a total of 80 samples in this testing dataset. The purpose of this test set is to compare the performances of selected models with results from chatGPT, as we only have access to the web-based interface of chatGPT, and the usage is limited.

EDOS Sexism Test Set is extracted from the EDOS dataset, containing only hate speech without reference counter-speech. This test set consists of all instances from the EDOS dataset that are labeled as sexist. The purposes of choosing a separate dataset as the second test set are (1) to test the model performances on texts written in different styles and (2) to evaluate the model performances on the topic of sexism. The EDOS dataset consists of sexist samples collected from Gab and Reddit, which are mostly written in a different style than the niche-sourced CONAN datasets, representing the ordinary use case of an SMP user more accurately.

Human Evaluation Test Set is a subset of the EDOS Sexism Test Set used for human evaluation, containing 10 randomly selected hate speech samples from the EDOS Sexism Test Set.

We excluded the Benchmark dataset from Qian due to its inconsistency in quality. The counter-speech responses in Qian's datasets are crowdsourced, often containing repeating and generic samples. These samples mostly refer to the community rule and state that violating such rules is not acceptable without confronting the hateful messages or providing factual information to counter the harmful intent.

The CONAN datasets are created via niche sourcing, where NGO operators who received training in constructing counter-speech were recruited. The expert-written counter-speech samples are, on average higher in quality. They are more relevant and engaging and often provide factual information to support their counterarguments. The knowledge-grounded Hate Countering Dataset is not included. Since this dataset is constructed using specifically chosen background knowledge, we have the concern that by including these samples, the trained model might hallucinate fake knowledge based on the training samples.

For training the topic classifier model, we excluded the CONAN dataset because of the imbalance in target groups it will cause. The CONAN dataset contains samples focused on Islamophobia, while the other two selected datasets are diverse in target groups.

Hate Speech	i joined gab to remind myself how retarded jew haters are. You wouldn't be typing on your abacus without them you retard.
Response	1. Using words that insult one group while defending another group doesn't come across as helpful. 2. You can make the same point more effectively without the use of hateful terminology. 3. Use of the r-word is unacceptable in our discourse as it demeans and insults people with mental disabilities.

Table 4.2.: An example from the Qian dataset (Gab) [60], where for one hate speech, three counter-speech responses are collected. Two of the three counter-speech responses are generic or refer simply to the community rules.

4.2.2. Data Preprocessing

For our experiment pipelines, we bring all samples into a uniform structure by applying several pre-processing techniques to clean and format each dataset. Firstly, we aim to bring each sample into the structure of single-turn hate- and counter-speech pair plus the targeted demographics. An example is shown in Table 4.3.

Hate Speech	You think these Muslim cockroaches have won? No, now is time for bloody retribution!
Counter Speech	What are they trying to win? Muslims in the UK aren't at war with anyone, the only thing creating conflict is Islamophobic rhetoric like this, seeking to create a divide.
Target	Muslims

Table 4.3.: An example from the training dataset in the desired format.

For the Custom Dataset, the required preprocessing is as follows:

Multi-target CONAN dataset is most suitable for our experiment, as it provides single-turn HS-CS pairs, each labeled with one or multiple target groups. We use this dataset as the basis of our Custom Dataset for training and fine-tuning further models.

CONAN dataset focuses on Islamophobia specifically. In addition to the target group ***Muslims***, each sample is labeled with sub-targets that signify the topic of the hate speech ("crimes", "culture", "rapism", "women", "terrorism", "generic", "islamization", "economics"). As most of the sub-targets don't fit into our target groups besides "women", we add the target group "Women" to the samples labeled with "women" and discard the other sub-targets. We included the English portion of the CONAN dataset.

DIALOCONAN dataset uses 6 of the 8 target groups as the Multi-target CONAN. Therefore, it doesn't need further adjustments. As we focus on a single-turn counter-speech generation scenario, we include only the first turn of the multi-turn samples in our training data. Though a dis-aggregate approach is commonly adopted, where each turn is separated into an independent single-turn instance, we decided to include only the first turn. The consideration is that the first turn of the conversation is always guaranteed to be self-contained with all the context information. While the following turns might also be self-contained, they are most likely referring to arguments from previous turns. Isolating these samples might reduce the data quality.

For all our datasets, we applied a simple data-cleaning pipeline on all hate speech and counter-speech texts to remove extra white spaces, line breaks, or consecutive punctuations. In addition, we replace any web link and emojis into unified tokens ***URL*** and ***$EMOJI$***.

4.2.3. Summary

All datasets used in our experiments are shown again in table 4.4.

Dataset	Source	Data Points
Custom Dataset	CONAN / Multi-target CONAN / DIALOCONAN	9540 + 2386 HS-CS pairs
Topic Classification Dataset	Multi-target CONAN / DIALOCONAN	11982 + 1332 CS
Small Custom Test Set	Custom Dataset (test split)	80 HS-CS pairs (10 per target group)
EDOS Sexism Test Set	EDOS dataset	4854 sexist HS
Human Evaluation Test Set	EDOS Sexism Test Set	10 sexist HS

Table 4.4.: Summary of all datasets used in our experiments. HS: Hate Speech, CS: Counter-Speech

4.2.4. Prompt Design

For using autoregressive language models like GPT-2, prompt design is proven to be important for generation quality. In our experiments, we use a simple prompt design to indicate the beginning of hate speech and counter-speech. For each HS-CS pair, we prepare the prompt as

[EOS] Hate-speech: [HS] Counter-speech: [CS] [EOS]

where EOS is the end-of-sentence token for marking sentence boundaries, HS is the content of hate speech, and CS the content of counter-speech. During inference time, we use the prompt

[EOS] Hate-speech: [HS] Counter-speech:

to let the model generate counter-speech for a given hate speech input.

4.3. Experimental Setup

The goal of our experiments is to evaluate the effectiveness of current approaches for generating counter-speech, especially against hate speech targeting women. Our experiments can be described in 5 stages:

(1) Compare Baseline Models: We train/fine-tune the selected baseline models using the Custom Dataset if necessary. Each model is then evaluated on the test split of the Custom Dataset, and the best-performing version of each baseline model is selected for further evaluation. For our experiments, we use a beam-search generation strategy with the number of beams fixed to 3.

(2) Train Target-Group-Aware Model: Using the best-performing model as the backbone, we implement the suggested target group embedding and fine-tune the model again using the same Custom Dataset. The performance of this target-group-aware model is evaluated on the Custom Dataset test split as other baseline models for comparison.

(3) Evaluate Performance on Sexist Dataset: The best-performing models and the target-group-aware model are evaluated on the EDOS Sexism Test Set. We aim to evaluate the performance of these models on sexist hate speech specifically to see whether the target group embedding improves the generation quality of the model.

(4) Compare Performance with chatGPT: In addition, we evaluate the same set of models on the Small Custom Test Set, where we also collect responses using two versions of chatGPT. As GPT3.5 and GPT4 achieve impressive results, we aim to investigate whether the approach of fine-tuning smaller pre-trained LMs is competitive.

(5) Conduct Human Evaluation: Lastly, we select the top-3 models on the sexist dataset and the two chatGPT models for the human evaluation.

4.3.1. Baseline Models

To examine the performances of currently available approaches for generating counter-speech, we included 5 recent methods for our experiment.

BART: The BART model, introduced by Lewis et al. [44], is popular for sequence-to-sequence text generation tasks due to the denoising autoencoder architecture and pre-training. In our work, we include the BART large-sized model with 12 encoder and decoder layers and 400M parameters in total.[1] We fine-tuned the model using our custom dataset.

GPT-2: The GPT-2 model is the successor model to the original GPT model, developed and released by OpenAI [62]. This large autoregressive language model is proven to be effective in text-generation tasks. We include the GPT-2 medium-sized model with 24 decoder layers and 355M parameters in total. Similar to the BART model, the gpt2-medium model is also fine-tuned using our custom dataset.

chatGPT: The recently-released conversational model chatGPT has drawn the attention of both researchers and the general public. The initially released version is built upon the GPT-3.5 architecture, which has shown significant improvements in performances across tasks in the NLP domain. chatGPT is further fine-tuned via RL from human

[1] `https://github.com/facebookresearch/fairseq/tree/main/examples/bart`

feedback, aiming to improve both the generated language quality while generating non-negative responses. A recent release also provided access to chatGPT with GPT-4 backbone, the newest and largest model in the GPT family. Due to the limited usage, we include chatGPT-3.5 and chatGPT-4.0,[2] tested on the small custom test set. For generating counter-speech responses using chatGPT, we use a fixed prompt design to give instructions:

"You are a UN agent who focuses on fighting online hate speech. Your task is to write responses, so-called counter speech, to online hate speech targeting different demographics. The responses should not be offensive, hateful, or toxic. The responses should actively fight against the hateful messages and bring counterarguments to the table. The aim is to bring positive perspectives, clarify misinformation, and be an active voice for the targeted demographics against hate. In the following, I will show you some online hate speech. Please create a counter-speech for each hate speech with a maximum length of 256 tokens.

[Hate Speech]"

GPS: The GPS pipeline, proposed by Zhu and Bhat [89], utilizes VAE for generating diverse counter-speech candidates and uses additional language models to prune and select the best fitting response for each hate speech input. We implemented the original pipeline[3] as described in the paper and trained the VAE using our custom dataset. We tested 14 candidate selection methods provided by Zhu and Bhat [89] on our custom dataset test split and selected the best-performing method for further testing on the EDOS sexism test set.

CounterGEDI: The CounterGEDI pipeline deploys GEDI models to guide model generation towards desired attributes [64]. In our experiments, we implemented the CounterGEDI pipeline as described in the original paper. A pretrained DialoGPT model is selected as the generation model and fine-tuned on our custom dataset. We adopted two GEDI models from the original paper for the attributes "***toxic***" and "***polite***", trained as described by Saha et al. [64]. We tested the generation model with and without GEDI influences on the custom dataset test split and selected the better-performing method for further testing.

4.4. Automatic Evaluation

In our work, we listed the most important attributes for generated counter-speech to be effective:

Language Quality: While human language is not always flawless and grammatical, correct sentences are generally a good starting point for constructing high-quality counter-speech. We assume that counter-speech should be structured correctly according to the

[2] 2023 May 24. Version
[3] `https://github.com/WanzhengZhu/GPS`

syntax and rules of the English language and use the **grammaticality** score to evaluate the language quality of any given counter-speech.

Toxicity: The counter-speech should be non-negative, as its purpose is to counter hateful messages without further escalating the conversation. Therefore, we want to measure whether a generated counter-speech is **toxic**, **offensive**, or **hateful**.

Valid as Counter-Speech: Naturally, a generated response should be a valid counter-speech. We define counter-speech as a counter-argument against hateful messages. It can, e.g., refute misinformation by providing accurate, evidence-based facts, promote empathy and understanding and humanize targeted demographics or victims, advocate for respectful, constructive conversation, or state consequences or impacts of hateful messages to prevent the hate from further spreading.

Diversity: Zhu and Bhat [89] suggest that diversity contributes significantly to the effectiveness of counter-speech. A set of counter-speech can be considered diverse if the samples are not largely commonplace and repetitive.

Relevancy We suggest that high-quality counter-speech should be relevant to the topic mentioned in the hateful messages. An effective counter-speech should address the main issue in the hateful messages and avoid using off-topic arguments (e.g., talk about discrimination against ethnical minorities while the hate speech promotes violence against women). It should also avoid using only generic responses that are less impactful (e.g., "You should be more understanding, and please don't use derogatory language.").

For each attribute, we selected automatic metrics to evaluate the generation responses.

4.4.1. Language Quality

For evaluating the **grammaticality** of the generated responses, we use language model classification scores. A roBERTa-based model introduced by Morris et al. [54] is implemented for evaluating how probable the given response is grammatically correct.

"textattack/roberta-base-CoLA" was fine-tuned on the Corpus of Linguistic Acceptability (CoLA) [78] subset of the glue dataset [76], a dataset with 10,657 English sentences labeled as grammatical or ungrammatical from linguistics publications. The CoLA classifier takes in an input sequence with a maximum length of 128 tokens and computes a Grammaticality score between 0 and 1, indicating the model's confidence of the input sequence being grammatical.

4.4.2. Toxicity

As suggested by Yeh et al. [83], using a combination of different models achieve better performance more consistently. Therefore, we include a **roBERTa-based toxicity classifier**, a **BERT-based toxicity classifier**, and a **BERT-based hate speech classifier** in our evaluation pipeline for assessing the responses' toxicity level.

's-nlp/roberta_toxicity_classifier': The roBERTa-based toxicity classifier is trained on the merge of the English parts of three Jigsaw datasets,[4,5,6] containing around 2 million examples. It takes in an input sequence and outputs a toxicity score between 0 and 1, with 1 indicating the sequence is toxic and vice versa.

'martin-ha/toxic-comment-model': The BERT-based toxicity classifier is trained on only the second Jigsaw dataset mentioned above (Unintended Bias in Toxicity Classification). Similarly, this model computes a toxicity score between 0 and 1 as the confidence of the sequence being toxic.

'Hate-speech-CNERG/bert-base-uncased-hatexplain': The BERT-based hate speech classifier is trained on the HateXplaine dataset[7] [50], a dataset containing 20K posts collected from Gab and Twitter, Each post labeled either hate, offensive or normal. The classifier computes three scores for a given sequence, one for **hatefulness**, one for **offensiveness**, and one for **normal**, indicating the probability of the given sequence belonging to the corresponding class.

In this category, we document three separate scores: **toxicity**, **offensiveness**, and **hatefulness**. Toxicity is jointly measured by the first two toxicity classifiers, while the "bert-base-uncased-hatexplain" classifier model is responsible for the offensiveness and hatefulness scores.

4.4.3. Valid as Counter-Speech

Since there is no existing counter-speech classification model, we fine-tuned a **BERT-based counter-speech classifier** for deciding whether a response is a valid counter-speech.

BERT-based Counter-Speech Classifier: The counter-speech classifier is fine-tuned on a counter-argument classifier[8] provided by ThinkCERCA, as counter-speech is in nature a type of counter-argument. We fine-tune the classifier using the custom dataset described in 4.2, using hate speech as negative samples and counter-speech as positive samples. The classifier computes the probability of a sequence being valid counter-speech, outputting a value between 0 and 1.

4.4.4. Diversity

For evaluating the diversity of a set of counter-speech, we implemented the n-gram Repetition Rate.

[4] `https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge`
[5] `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification`
[6] `https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification`
[7] `https://github.com/hate-alert/HateXplain`
[8] ThinkCERCA/counterargument_hugging

Repetition Rate (RR): RR [12] evaluates the repetitiveness of a collection of texts. This metric computes the rate of non-singleton n-gram types it contains and indicates the capability of the language model to produce responses with high lexical diversity. A high RR score means more repetitive generation results, indicating a less diverse language generation ability. We follow the recommendation by Cettolo et al. [16] and set $n = 4$

4.4.5. Relevancy

A common approach for measuring the relevancy of a given response is to calculate the similarity between the response and the label or, in some cases, the response and the input sequence. Metrics like RUBER [70] and BERT-RUBER [32] utilize language models to compute similarities. Following this approach, we proposed to use (1) **Universal Sentence Encoders** for computing cosine similarity between the generated response and the reference response and (2) **Question-Answering Sentence Encoders** for computing cosine similarity between the generated response and the input hate sequence.

Universal Sentence Encoders are commonly trained to capture semantic information of a given sentence or short paragraph. The produced contextual representation of two sentences with similar meanings should ideally be similar as well. This characteristic makes them attractive for computing the similarity between a generated response and the reference response if we assume that the generated responses should be semantically similar to the reference. As suggested by Yeh et al. [83], using a combination of different models achieve better performance more consistently. Therefore, we included three universal sentence encoders:

'all-MiniLM-L6-v2': This sentence-transformers model maps input sequences to a 384 dimensional dense representation. It is fine-tuned based on the *nreimers/MiniLM-L6-H384-uncased* model using a contrastive learning objective on a 1B sentence pairs dataset.[9]

'all-mpnet-base-v2': This sentence-transformers model maps input sequences to a 768 dimensional dense representation. It is fine-tuned the same way as *'all-MiniLM-L6-v2'*, but based on the *microsoft/mpnet-base* model.

'LaBSE': LaBSE (Language-agnostic BERT sentence embedding model) is a sentence encoder that provides support for 109 languages. This model is trained and optimized to produce similar representations for bilingual sentence pairs that are translations of each other. It produces similar contextual embeddings for sentences containing similar content, which can be used for semantic similarity comparison.

Question-Answering Sentence Encoders are similar to universal sentence encoders but trained on datasets containing question-answer pairs. In addition to capturing semantic information, they are trained to produce similar representations for question-answer pairs

[9] `https://discuss.huggingface.co/t/train-the-best-sentence-embedding-model-ever-with-1b-training-pairs/7354`

that are highly relevant. By computing similarity scores, such models can find relevant sequence candidates for given passages. We included two question-answering sentence encoders for computing similarity scores between the input hate sequence and generated counter-speech:

'multi-qa-MiniLM-L6-cos-v1': This sentence-transformers model simultaneously maps a query sequence and a series of candidate sequences to a 384 dimensional dense representation per sequence. The cosine similarity between the query and a candidate representation indicates their relevancy. It is fine-tuned based on the *nreimers/MiniLM-L6-H384-uncased* model using a contrastive learning objective on 215M question-answer pairs from diverse sources under the same project as *'all-MiniLM-L6-v2'*.

'multi-qa-distilbert-cos-v1': Similar to *'multi-qa-MiniLM-L6-cos-v1'*, this model maps query and candidate sequences into 768 dimensional representations. It is trained on the same setting as *'multi-qa-MiniLM-L6-cos-v1'* based on the *'distilbert-base-uncased'* model.

Since the cosine similarity range is originally $\{-1, 1\}$, we normalize the score to $\{0, 1\}$ for a uniform scoring system across attributes. As discussed in 2.4, similarity-based metrics assume an overlap between the response and a reference sequence, whether the label or the input. This is not always true, and a high similarity score does not directly reflect whether the response is addressing the same topic as the hate speech. To address this issue, we introduced a **roBERTa-based topic classifier** for classifying whether a response is relevant regarding specific topics.

Sexism-Topic Classifier is fine-tuned based on the *'cardiffnlp/tweet-topic-21-multi'* model, a multi-label topic classifier trained to classify tweets in 14 topics. We fine-tuned the model with a new classification head for our 8 topics using the topic classification dataset.

4.4.6. Summary

Besides the similarity-based metrics and Repetition Rate, all other metrics are classifier-based. We decided on a hard classification scoring, meaning that each sample is either scored with 1 as positive or 0 as negative. For each classifier model, we set a threshold ϵ. When a classifier model outputs a probability $p > \epsilon$, the input sequence receives a score

$$s = \begin{cases} 1, & \text{if the classifier attribute is positive} \\ 0, & \text{otherwise} \end{cases}$$

E.g., the grammaticality classifier classifies the attribute "*grammatical*", which is a positive attribute. Therefore, a probability above the threshold would give the input sequence a score of 1. On the other hand, the toxicity classifier classifies the attribute "*toxicity*", which is negative. An input sequence with a probability above the threshold would receive a score of 0. If an attribute is evaluated via multiple classifier models, the final score for each input sequence is calculated as the average of all scores from every individual model.

For a set of counter-speech $C = \{c_1, ..., c_n\}$ generated by a model M, and each sample receives a score $S = s_1, ..., s_n$, we can calculate a model-wise arithmetic mean score S_M for each attribute as the average score over all counter-speech instances:

$$S_M = average(S) \tag{4.1}$$
$$= \frac{1}{n}\sum_{i=1}^{n} s_i \tag{4.2}$$

Attribute	Classifier/s	Classifier-/Positive-Attribute	Threshold
Langauge Quality	textattack/roberta-base-CoLA	grammatical/grammatical	0.5
Toxicity	s-nlp/roberta_toxicity_classifier martin-ha/toxic-comment-model	toxic/non-toxic	0.5
Hatefulness Offensiveness	Hate-speech-CNERG/bert-base-uncased-hatexplain	hateful/non-hateful offensive/non-offensive	0.5
Valid as Counter-Speech	BERT-based Counter-Speech Classifier	valid as counter-speech/ valid counter-speech	0.95
Relevancy	roBERTa-based Topic Classifier	topics (categorical)	0.5

Table 4.5.: Summary of classifier-based evaluation metrics

Attribute	Metric / Sentence Encoder
Diversity	Repetition Rate
Label Similarity	all-MiniLM-L6-v2 / all-mpnet-base-v2 / LaBSE
Context Similarity	multi-qa-MiniLM-L6-cos-v1 / multi-qa-distilbert-cos-v1

Table 4.6.: Summary of non-classifier-based evaluation metrics

4.5. Human Evaluation

To verify the results from the automatic evaluation, we prepared a human assessment framework for evaluating the same attributes of each model via human judgment. In particular, our aim is to evaluate three objective attributes that are inherently complex for automatic metrics to evaluate accurately, namely *Toxicity*, *Relevancy*, and *Validity as Counter Speech*.

Design: We selected 10 hate speech samples randomly from the EDOS Sexism Test Set as our Human Evaluation Test Set. In combination with counter-speech generated by the 5 selected models, we end up with a total of 50 HS-CS pairs for the human evaluation. Each HS-CS pair is rated in regard to the three attributes: *Toxicity*, *Relevancy*, and *Validity as Counter-Speech*.

Due to capacity constraints, we adopted a streamlined bipolar evaluation scheme, where each attribute is scored as either "Yes" or "No". This approach mirrors the classifier-based automatic metrics, where each sample is classified into one of two categories, labeled "0" or "1". This design is more efficient than employing a finer scale, as it allows for easier recruitment and reduces the cognitive workload on the evaluations. In the interest of elucidating the process, an example from the evaluation sheet is shown in figure 4.2. The complete evaluation sheet is attached in Appendix A.

Hate_Speech	Counter_Speech	Toxic/ Offensive	Relevancy	Acceptable as Counter-Speech
Slapping a bitch is quite rewarding. He needs to learn how to punch though. If he would have thrown that straight right like a man instead of a bitch, she would have been eating the pavement.	Did you know that punching someone in the face with a closed fist is considered a crime in the US? In fact, it is a civil violation punishable by up to a year in jail and a fine of up to $1,000.00. If you want to learn more about civil rights, check out the article on the American Civil Liberties Union about the history of racial discrimination in the United States. Also, if you think that a man can't be a feminist, then you need to check your privilege. You are a man, and you are allowed to be a misogynistic man.			

Figure 4.2.: An example from the human evaluation sheet.

Participants: 6 University students are recruited from diverse study programs and degrees, ranging from undergraduate to Ph.D. candidates. The participants' knowledge of counter-speech creation or their engagements in online hate is not surveyed. However, we assumed that most participants have not received professional training in detecting hate speech nor in constructing counter-speech according to specific guidelines.

Procedure: The evaluation process is conducted online, utilizing Google Sheets as the questionnaire. We arranged an information page with explicit instructions and clarification of relevant notations for the participants. Additionally, we provided simplified examples that serve as basic guidelines for common scenarios.

Each participant is expected to complete the evaluation sheet independently. Upon completion, participants are instructed to return the completed sheet. The independent completion minimizes potential biases that could result from collaborative completion, thereby ensuring more robust and reliable data.

Instructions: Participants are instructed to evaluate each HS-CS pair in regard to the three attributes. To facilitate ease of response and uniformity in data collection, participants are provided with a predefined drop-down menu from which they can select either "Yes" or "No".

Measurements: Similar to the automatic evaluation, we ascribe a numeric value of 1 to a 'Yes' response and 0 to a 'No' response. Firstly, we calculate the average attribute score for each model across the 10 samples, obtaining a participant-wise average score. The individual participant score is then averaged across the entire participant pool, resulting in a model-wise average score for each attribute.

5. Results

In this Chapter, we present the experimental results that were obtained in our study. The chapter is structured into two sections, focusing on automatic evaluation results and human evaluation results, respectively. Section 5.1 begins with the performance leaderboards of the baseline models. We provide a comprehensive analysis of the scoring and highlight the most significant findings that emerged from the evaluation process. Moving on to Section 5.2, we present the leaderboard for the human evaluation results. This section offers a detailed examination of the annotator agreement situation, providing additional insights into the assessments made by the human annotators.

5.1. Automatic Evaluation

The results from the automatic evaluation are shown in table 5.1, 5.2, and 5.3. Notably, the label references attained the highest scores in both the Custom Dataset and the Small Custom Test Set. This outcome is significant as it implies that the counter-speech responses crafted by experts are of superior quality overall. Therefore, it suggests that our evaluation pipeline is reliable and, to some extent, reflects human judgment.

The evaluation results consistently show that large pre-trained LMs outperform other approaches in terms of the arithmetic mean score. The advantage of large pre-trained LMs, particularly in language quality, is evident from significantly higher CoLA scores compared to VAE-based approaches.

When comparing automated approaches with human performance, the largest discrepancy exists in the scoring of the Repetition Rate. As expected, automated approaches struggle to generate diverse responses compared to humans. The GPS model with Universal Sentence Encoder matching demonstrates the best performance in terms of diversity, as it is specifically optimized for generating diverse responses.

In the case of chatGPT, both versions of the model exhibit the ability to produce grammatically correct, non-toxic, non-hateful, and non-offensive counter-speech responses. However, they face challenges in generating diverse responses, as indicated by the lower Repetition Rate scores. Upon further investigation, we attribute this limitation to the prompt design employed when interacting with chatGPT. The repeated use of the same prompt often leads the model to generate responses with similar paragraph structures and, at times, even identical sentences for hate speech on similar topics.

The evaluation results from the EDOS Sexism Test Set provide interesting insights. Firstly, we observe that the selected models perform reasonably well on unseen data with different styles. The CONAN dataset, which consists of expert-written samples including hate speech,

Model Name	**CoLA**	**TOX**	**Hate**	**OFF**	**L Sim**	**C Sim**	**VaCS**	**RR**	**F1**	**AM**
Human	0.937	0.955	1.000	**0.997**	-	0.751	**0.980**	**0.861**	0.885	**0.929**
gpt2-medium	**0.964**	**0.967**	0.999	0.995	0.716	0.798	**0.967**	0.383	0.893	**0.853**
bart-large	0.948	0.941	1.000	0.995	0.723	0.786	0.953	0.466	0.839	0.850
target-aware gpt2-medium	0.958	0.946	1.000	**0.996**	0.706	0.784	0.946	0.419	0.880	0.848
GPS [USE-QA-SIM]	0.793	0.887	1.000	0.986	0.728	0.838	0.946	0.500	**0.944**	0.847
GPS [USE-QA-MAP]	0.812	0.916	1.000	0.990	**0.730**	0.821	0.949	0.535	0.861	0.846
GPS [USE-LARGE-MAP]	0.807	0.922	1.000	0.988	0.718	0.806	0.923	0.558	0.809	0.837
GPS [USE-MAP]	0.789	0.922	1.000	0.992	0.713	0.799	0.922	**0.564**	0.803	0.834
GPS [BERT-SMALL-MAP]	0.818	0.911	1.000	0.981	0.714	0.801	0.927	0.534	0.805	0.832
GPS [ELMO-MAP]	0.829	0.922	1.000	0.989	0.709	0.789	0.949	0.505	0.770	0.829
GPS [BERT-LARGE-MAP]	0.768	0.900	1.000	0.988	0.708	0.796	0.895	0.562	0.804	0.825
GPS [USE-LARGE-SIM]	0.734	0.847	1.000	0.973	0.723	**0.836**	0.868	0.483	0.918	0.820
GPS [USE-SIM]	0.726	0.841	1.000	0.972	0.718	0.826	0.865	0.506	0.924	0.820
GPS [BERT-LARGE-SIM]	0.883	0.865	1.000	0.985	0.702	0.792	0.876	0.481	0.795	0.820
GPS [ELMO-SIM]	0.842	0.904	1.000	0.995	0.696	0.777	0.955	0.390	0.780	0.815
GPS [BERT-SMALL-SIM]	0.821	0.844	1.000	0.962	0.699	0.794	0.844	0.474	0.791	0.803
GPS [TF-IDF]	0.746	0.882	1.000	0.981	0.684	0.760	0.879	0.536	0.610	0.786
CounterGeDI [p + nt]	0.919	0.837	0.996	**0.996**	0.668	0.733	0.759	0.418	0.682	0.778
CounterGeDI	0.861	0.858	1.000	1.000	0.679	0.761	0.771	0.309	0.595	0.759
GPS [BM25]	0.680	0.944	1.000	0.986	0.653	0.702	0.936	0.398	0.479	0.753

Table 5.1.: Automatic evaluation leaderboard on the Custom Dataset test split, sorted based on AM score. The highest score in each attribute is marked in bold. If the best score is human performance, the second-best score is marked as well. Notations see Table 5.2.

and the EDOS dataset, collected from Gab and Reddit, exhibit distinct characteristics. Despite this variation, the models show a relatively consistent performance across the datasets.

The most significant drop in performance across models is observed in the target-demographic classifier's F1 score. In the training dataset, most models achieved high relevancy scores, averaging above 0.8. However, on the Sexism Test Set, the average relevancy score decreases to around 0.5, except for the target-demographic aware GPT2 model. By providing target-demographic information to the model through target-demographic embeddings, we achieved a significant improvement in the relevancy metric. This suggests the effectiveness of our approach in assisting the model in generating more relevant responses.

However, compared to the GPT2 model without target-demographic embeddings, we observe a small but consistent decrease in performance in other aspects. This phenomenon requires further investigation. Our speculation is twofold: Firstly, adding new tokens to the vocabulary may have affected the fine-tuning process, leading to some degree of "forgetting" of the pre-training. Secondly, in autoregressive training, the target-demographic tokens are also subject to prediction. Ideally, the target-demographic tokens should only be present in the input sequence, providing information, but never generated in the output sequence. This is, for instance, the case in fine-tuning sequence-to-sequence models, where target-demographic

Model_Name	CoLA	TOX	Hate	OFF	L Sim	C Sim	VaCS	RR	F1	AM
Human	0.963	0.956	1.000	1.000	**1.000**	0.768	0.988	**0.995**	0.868	**0.949**
ChatGPT-4.0	**1.000**	**1.000**	1.000	1.000	0.746	0.812	**1.000**	0.875	**0.888**	**0.924**
bart-large	0.963	0.938	1.000	1.000	**0.766**	0.799	0.950	0.905	0.830	0.906
gpt2-medium	0.988	0.962	1.000	1.000	0.746	0.809	0.963	0.859	0.807	0.904
ChatGPT-3.5	**1.000**	**1.000**	1.000	1.000	0.723	0.758	**1.000**	0.742	0.863	0.898
GPS [USE-QA-SIM]	0.813	0.875	1.000	0.963	0.761	**0.839**	0.963	**0.979**	**0.888**	0.898
target-aware gpt2-medium	0.975	0.931	1.000	1.000	0.728	0.783	0.888	0.911	0.792	0.890
CounterGeDI [p + nt]	0.888	0.800	0.988	1.000	0.669	0.710	0.725	0.877	0.403	0.784

Table 5.2.: Automatic evaluation leaderboard on the Small Custom Test Set, sorted based on AM score. The highest score in each attribute is marked in bold. If the best score is human performance, the second-best score is marked as well. **[p + nt]**: guided with politeness and non-toxic GEDIs; **TOX**: Toxicity; **Hate**: Hatefulness; **OFF**: Offensiveness; **L Sim**: Label Similarity; **C Sim**: Context Similarity; **VaCS**: Validity as Counter-Speech; **RR**: Repetition Rate; **F1**: target-demographic classifier F1 score; **AM**: Arithmetic Mean Score.

tokens never appear in the target sequence. However, during autoregressive training, the model assigns a certain probability to these tokens, occasionally generating them in the output sequence.

To address these challenges, further research is needed to mitigate the decrease in performance observed in aspects other than relevancy when incorporating target-group embeddings. Techniques such as token masking or alternative training strategies could be explored to minimize the impact on fine-tuning and ensure that target-group tokens are appropriately handled during autoregressive generation.

Model_Name	CoLA	TOX	Hate	OFF	C Sim	VaCS	RR	F1	AM
bart-large	**0.957**	0.863	0.995	0.960	**0.747**	0.837	**0.696**	0.550	**0.826**
target-aware gpt2-medium	0.930	0.815	0.999	0.975	0.689	0.857	0.518	**0.747**	0.816
gpt2-medium	0.956	0.903	0.998	0.984	0.705	0.933	0.520	0.520	0.815
GPS [USE-QA-SIM]	0.733	**0.927**	**1.000**	0.981	0.683	**0.937**	0.434	0.587	0.785
CounterGeDI [p + nt]	0.914	0.793	0.999	**0.990**	0.611	0.576	0.528	0.240	0.706

Table 5.3.: Automatic evaluation leaderboard on the EDOS Sexism Test Set, sorted based on AM score. The highest score in each attribute is marked in bold. If the best score is human performance, the second-best score is marked as well. The label similarity is excluded since no reference label is available for this dataset. Notations see Table 5.2.

5.2. Human Evaluation

The results of the human evaluation are presented in Table 5.4. As the leaderboard differs from the automatic evaluation results, a closer examination of the results, along with the analysis of annotator agreement depicted in Figure 5.1, is warranted.

Model Name	TOX	VaCS	REL	AM
chatGPT 4	**1.000**	**0.950**	**0.933**	**0.961**
chatGPT 3.5	1.000	0.683	0.833	0.839
gpt2-medium	0.817	0.533	0.550	0.633
bart-large	0.650	0.300	0.450	0.467
target-aware gpt2-medium	0.600	0.350	0.433	0.461

Table 5.4.: Human evaluation leaderboard, sorted based on AM score. The highest score in each attribute is marked in bold. **TOX**: Toxicity; **VaCS**: Validity as Counter-Speech; **REL**: Relevancy; **AM**: Arithmetic Mean.

Based on the leaderboard results, both versions of chatGPT have performed exceptionally well. As indicated by the automatic evaluation results, chatGPT is capable of generating fluent, grammatical, and highly relevant counter-speech in response to the given hate speech sequences. This is further supported by the annotator-agreement situation, where for chatGPT using the GPT-4 backbone, most annotators agreed on similar scoring within a narrow score range of 0.2. The chatGPT model using the GPT-3.5 backbone also exhibited more consistent scoring in terms of toxicity and relevancy. However, for the other models, there was strong disagreement among annotators, posing challenges in interpreting the results.

From the current statistics, toxicity scoring appears to be the most consistent attribute among the three. This is likely due to the more widely accepted definition of toxic or offensive speech. However, determining the validity of counter-speech and assessing relevancy is more challenging, as the definitions of counter-speech and relevancy can be subjective and vary among individuals. Despite providing a few examples, annotators provided feedback stating that some hate speech and counter-speech samples were difficult to understand, and the decision criteria were vague in most cases, leading to confusion in making judgments. This indicates that further improvement is needed in the instructions and explanations provided in the human evaluation pipeline. Additionally, a larger annotator pool and sample size are necessary to obtain more informative evaluation results.

While it is challenging to draw individual conclusions about the performance of each model, it is evident that extremely large models like GPT-4 can effectively extract information from the input sequence and generate high-quality responses that receive approval from most annotators. This suggests the potential of such models in producing relevant and effective counter-speech, although improvements are required in the evaluation process to provide more reliable and robust assessments.

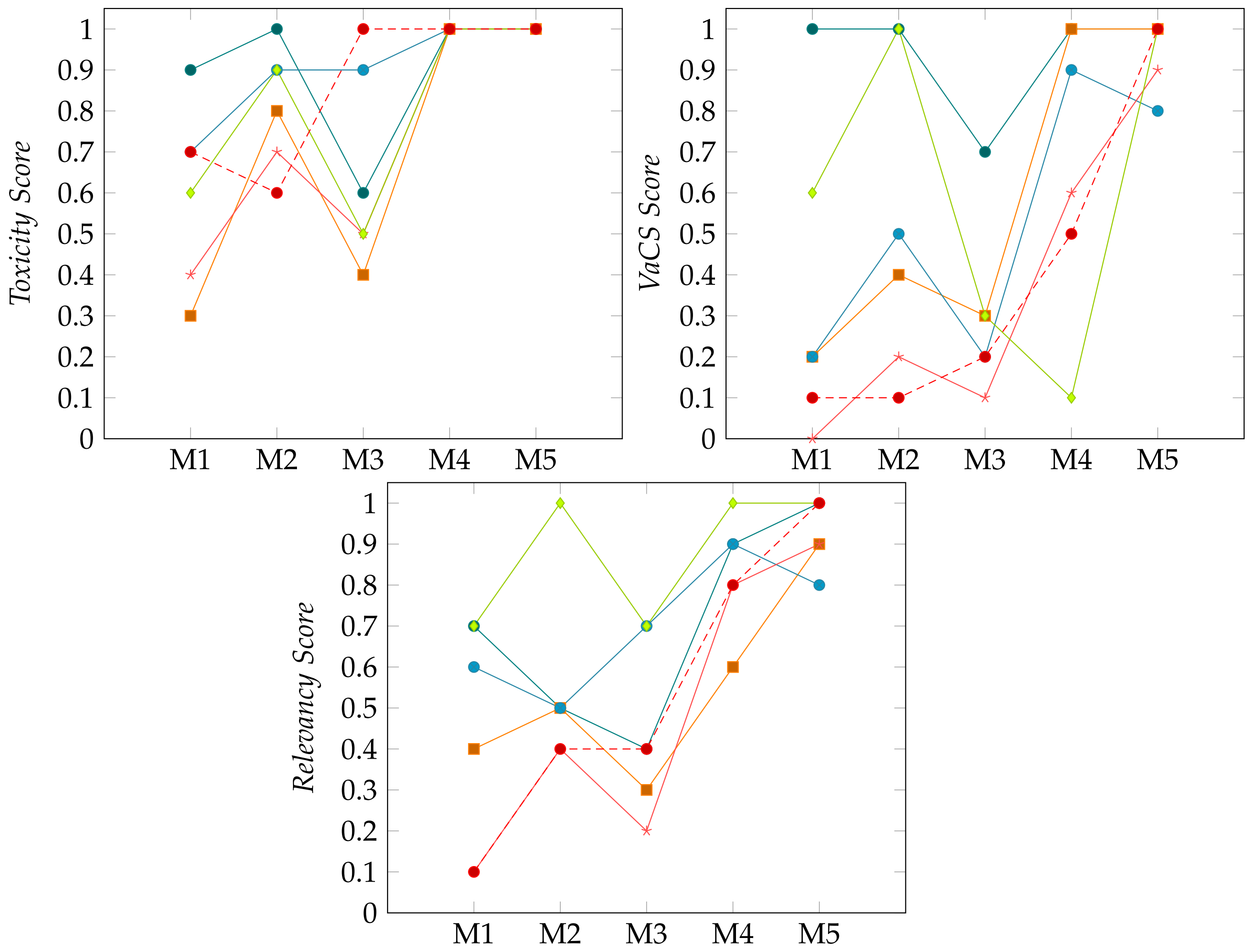

Figure 5.1.: Annotator-wise analysis of the human evaluation results for each attribute. **Top-left:** Toxicity Score; **Top-right:** Validity as Counter-Speech Score; **Bottom:** Relevancy Score; Each colored line indicates one participant's score across the 5 models. **M1**: target-aware gpt2-medium; **M2**: gpt2-medium; **M3**: bart-large; **M4**: chatGPT 3.5; **M5**: chatGPT 4;

6. Conclusion

The ubiquity of online hatred in the information era, amplified by the accelerated digitalization of day-to-day activities, poses significant challenges as it rapidly spreads through social media platforms. In response, our research has focused on harnessing recent advancements in language modeling to explore automated approaches for generating counter-speech. Additionally, we have developed evaluation frameworks to assess the quality and effectiveness of the generated counter-speech. Furthermore, we have introduced a generative framework that incorporates target-demographic information through special tokens, aiming to produce counter-speech that is more contextually relevant.

In the concluding chapter of our study, we reflect on the research questions posed in Chapter 1:

Q1: *"To what extent do current techniques effectively address hate speech?"*

Through our investigations, we have demonstrated that current techniques, particularly the use of large pre-trained language models such as GPT-4, hold promise in effectively addressing hate speech. While access to such models may be limited and the training process resource-intensive, we have also explored the potential of fine-tuning smaller pre-trained models like BART and GPT-2 in automated counter-speech generation pipelines. These pipelines have showcased the ability to generate contextually relevant responses that are fluent, grammatically correct, and aligned with the goal of combating hate speech.

Q2: *"What strategies can be employed to generate counter-speech that is more contextually relevant in combating gender-specific hate speech?"*

Our research has specifically investigated strategies to enhance the contextual relevancy of counter-speech, with a focus on addressing gender-specific hate speech. By incorporating target-demographic information through special tokens, we have demonstrated improvements in generating counter-speech that is more relevant to the specific target group.

6.1. Limitations and Future Works

While this research has demonstrated the feasibility of generating contextually relevant counter-speech responses using automated pipelines, it is important to acknowledge several limitations and areas for future investigation.

Firstly, evaluating the quality of dialogue responses poses a challenge due to the absence of clear criteria and robust evaluation techniques. A high-quality counter-speech should ideally be effective, credible, and convincing [21, 28, 64, 89]. In our proposed evaluation pipeline, we have incorporated several essential factors that contribute to high-quality counter-speech, including fluency and grammatical accuracy, politeness and non-toxicity, validity as counterarguments, and contextual relevancy. However, the effectiveness and credibility of the generated responses are not directly measured. While our evaluation pipeline offers more informative insights compared to traditional similarity-based metrics, further refinement is necessary. In the following sections, we will highlight several improvements that we intend to implement beyond the scope of this research:

Measure Credibility: Although our research did not specifically address the detection or mitigation of content hallucination, recent studies have investigated methodologies for identifying hallucinated content in computer-generated texts through the fine-tuning of classification models [88]. By incorporating these models into future evaluation pipelines, it would be possible to directly measure the credibility of the generated responses as one of their attributes. Furthermore, active efforts have been dedicated to developing knowledge-grounded generation methods [22], which utilize external knowledge sources to ensure the accuracy and factual basis of the generated counter-speech.

Similarity-based Metrics: In our research, we utilized two similarity-based metrics to evaluate the relevance of the generated responses. However, we acknowledge that the interpretability and informativeness of these metrics were questioned during the assessment of the automatic evaluation results. While these metrics provide an indication of the model's performance, it was challenging to determine precisely what aspects they measure. Recognizing this limitation, it would be advantageous to explore alternative evaluation metrics that provide clearer insights and align more closely with the specific aspects of counter-speech generation.

Relevancy Classifier: In addition to the similarity-based metrics, we introduced a target demographic classification model to assess the relevancy between hate speech and counter-speech pairs. We believe that this classifier-based metric can provide more informative results, as evidenced by the automatic evaluation. However, the human evaluation scores, despite being challenging to interpret, have raised questions about the effectiveness of this metric. Therefore, we propose several adaptations to further improve its performance.

To train the classifier, we utilized the labels of hate speech as the counter-speech label. While this approach is generally acceptable, it may result in the inclusion of numerous generic responses that are labeled as relevant, potentially impacting the overall performance of the model. To address this concern, further investigation into alternative training methods could be beneficial. For instance, using more specific labels or employing a larger and more diverse dataset could enhance the accuracy of the classifier.

Moreover, the human evaluation regarding relevancy exhibited strong disagreement among annotators. We posit that this issue may stem from a more fine-grained definition of "relevancy" for humans. In our classifier model, a pair of hate speech and counter-speech is deemed contextually relevant if they pertain to the same pre-defined topics within eight target demographics. However, during the human evaluation, annotators reported instances where they considered a response relevant to the same topic but not directly relevant to the original hate speech content. This indicates a failure on our part to provide clear training and guidelines for annotators. Additionally, the coarse definition of the eight target demographics may not be the ideal approach. To address these concerns, we propose the construction of larger and more diverse datasets comprising single-turn dialogues labeled as relevant or irrelevant. Utilizing such datasets, we can directly train a relevancy classifier that captures a broader understanding of relevancy in counter-speech generation.

Human Evaluation: The unexpected but informative situation of strong annotator disagreement has shed light on important considerations. It has become evident that there is a need for enhanced annotator training and larger sample size to ensure more robust results. In our current evaluation, having only six annotators assess 50 samples proved insufficient, as individual outliers could disproportionately impact the overall outcomes. Furthermore, the absence of comprehensive training or guidelines for annotators resulted in differing interpretations of decision boundaries.

To address these issues, we recommend conducting further human evaluations using a larger number of hate speech and counter-speech pairs. In order to ensure a more uniform and professional understanding of counter-speech construction, it would be beneficial to recruit experienced NGO operators or trainers who have undergone specific training in this domain. Their expertise can contribute to a more consistent assessment process and provide valuable insights into the effectiveness and relevancy of the generated counter-speech responses. By incorporating their perspectives and leveraging their expertise, we can achieve more reliable and meaningful human evaluation results.

Secondly, while our research demonstrated that providing additional information to the models through new tokens can enhance the relevancy of the generated responses, the distinction between using special tokens versus conveying the same information through natural language is not clear. As suggested by Ashida and Komachi [4] and other studies, prompt design can significantly impact the performance of models even without fine-tuning. Exploring diverse approaches to incorporate supplementary information beyond the input sequence and examining how they influence the quality of the generated counter-speech could provide valuable insights for future research.

Likewise, incorporating other features into the models could also prove beneficial in generating more effective counter-speech. As demonstrated in our work, simply providing the model with target-demographic information resulted in more relevant responses that addressed the specified target demographic. Taking a broader perspective, this approach can be considered a form of controlled generation, where the generation process is guided by added features. As pointed out by Chung [21], personalized persuasion techniques [77] that leverage the hater's user profile information can be highly effective. Building upon this concept, training generation models that can tailor the generated responses to the specific hate speech instances and the corresponding haters could further enhance the relevancy and effectiveness of the counter-speech.

Another intriguing direction arises when considering the scenario of generation models assisting NGO operators in constructing counter-speech. In such cases, NGO operators may wish to approach different instances of hate speech using varying tones or emotions. Incorporating such information into the training data can be immensely valuable in accelerating the workflow and improving the quality of the generated counter-speech. This approach could empower NGO operators by providing them with more nuanced and customized tools for addressing hate speech in a manner that aligns with their communication strategies and objectives.

Lastly, it is important to acknowledge that the choice of generation strategy can have a significant impact on the quality of the generated responses, as suggested by relevant research. During the testing of our proposed pipelines, we conducted experiments generating multiple responses, where the model selects the top n results for each hate speech instance. Due to the scope of this research, this aspect was not included in the final experimental setup. However, it is noteworthy that beyond the first best-scoring response, we observed a considerable decrease in language quality and an increase in toxicity. In our experiments, we employed beam search with a fixed number of beams set at 3, retaining only the best result.

Nevertheless, we did not explore alternative generation strategies or conduct a comprehensive analysis of how different strategies might impact the quality of the generated counter-speech. Future research could delve into the influence of various generation strategies, such as beam search with diversity-promoting strategies, on the effectiveness of counter-speech generation. By systematically examining and comparing different approaches, we can gain a better understanding of their strengths and weaknesses and identify strategies that consistently produce high-quality and contextually relevant counter-speech responses.

In conclusion, while this research makes advancements in the development and evaluation of counter-speech generation models to combat online hate speech, it is important to acknowledge the aforementioned limitations. Addressing these limitations in future studies will contribute to the refinement and effectiveness of counter-speech generation techniques.

A. Human Evaluation Sheet

Questionnaire for Evaluating Language Model Generation Results for Countering Online Sexist Hate

Instruction

- Please find the question sheet (Human_Evaluation_Sheet) using the menu at the bottom.

- During this questionnaire, you will be given 10 hate speech samples. For each hate speech, 5 counter speech are given, making in total 50 counter speech responses.

- For each counter speech response, three YES/NO questions are asked for evaluating three attributes of quality (**Toxicity**, **Relevancy** and **Acceptance as Counter Speech**).

- Each cell is pre-defined as a **drop-down menu**, where you can simply choose YES or NO from the list items. Please make sure to fill out all the cells.

- Once you're satistied with your answers, you can send back the questionnaire by simply sharing the file with yenyu.chang@hotmail.com or save the file as an excel file and send it via any means.

- Some examples as reference are provided in an extra sheet behind the actual question sheet, take a look if you can't decide what to do.

Warning

this project contains content that may be offensive or upsetting. These material are collected from the internet for the purpose of developing counter measures. These material are not produced by the authors of this project and do not represent the value of the authors.

Notion Explaination

Toxic/Offensive: A response is considered as toxic or offensive, if it contains contents that can be harmful, degrading, or hurtful towards others. Examples for toxi or offensive speech are:

- Hate Speech: This includes language that discriminates, stereotypes, marginalizes, or stigmatizes individuals or groups based on attributes such as their race, religio ethnic origin, sexual orientation, disability, or gender.

- Harassment: This can involve threats, intimidation, or repeated unwanted contact or attention.

- Disinformation: Spreading false or misleading information to cause harm or confusion.

- Personal Attacks or Insults: Directly attacking someone's character, appearance, or other personal attributes in a harmful manner.

- Profanity or Obscene Language: Using crude, vulgar, or offensive language.

Relevancy: A response is considered relevant, if it revolves generally around the same topic/target as the hate speech, and tackles the hateful messages in the origina hate speech.

- To judge the relevancy of sample, you can imagine putting the response to different hate speech and see if it can also fit. A general counter speech such as "Please do not use derogatory language and be understanding." is in most cases not specific enough to be considered relevant.

Acceptance as Counter Speech: A response is considered a counter-speech, if it's a counter argument against the original hate speech. It can:

- refute misinformation by providing accurate, evidence-based facts.

- promote empathy and understanding, humanize targeted demographics/victims.

- advocate for respectful, constructive conversation

- state concequences/impacts of hateful messages, prevents the hate from further spreading.

Hate_Speech	Counter_Speech	Toxic/Offensive	Relevancy	Acceptable as Counter-Speech	T Count	R Count	CS Count
But the law counts the teen girls that run away to look for sex work. Imo those arent missing kids or sex trafficking victims.	The law does not specifically prohibit the enforcement of this rule. In fact, it is a general rule tha does not apply to minors. specially when it comes to the exploitation of vulnerable people like minors (link to article about the law) If you really wanted to know more about this topic, you should talk to the people who actually enforce the law. Sexual exploitation of children is a serious crime that shoul be stopped. But it is not a crime only to be aware of it. It is also a crime to fail to prevent it.				0	0	0
	The law counts only those cases where a woman is forced to engage in sex work, not those where a man conspires to exploit her.				0	0	0
	It's crucial to approach the topic of missing teenagers and sex work with empathy and understanding. While some teenagers may run away and engage in sex work voluntarily, it's important to recognize that they may face vulnerabilities and exploitation. Let's focus on providing support and resources to protect these young individuals and address the underlying issues they may be facing. #SupportVulnerableYouth				0	0	0
	It's important to approach this sensitive topic with empathy. Many "runaway" teens are dealing with complex issues like abuse or neglect at home, and they may see no other options. It's not their choice to be in a situation that puts them at risk. They are indeed victims and need our understanding and help, not judgment. Let's focus on creating safe environments for all children.				0	0	0
	If you really cared about sex workers, you would know that they are just as scared of their employers as anyone else. They are forced to work in dangerous and unsafe conditions just to make a living. And they are often the only ones able to escape. If you really wanted to help them, you should start by supporting them. You can do it by supporting the sex workers' rights movement, or by donating to the victims of human trafficking. Homophobia and transphobia are not about sex work, they are about the oppression of women.				0	0	0
All her friends know divorce attorneys. Roastie	She has friends who are divorce lawyers, and they all know that divorce is a very serious matter. If you have friends who know this, then you are in the minority. If not, then maybe you should talk to them about it. And if you still don't feel good about the idea of divorce, then perhaps it is time to talk to a mental health professional. You might find that you are not the only one suffering from post-traumatic stress disorder after divorce. Counterspeech: What do you mean by friends?						
	That sounds like a stereotype. What do you mean by that?						
	I understand that you may have certain opinions or frustrations, but it's important to maintain respectful and constructive dialogue. Using derogatory terms or engaging in personal attacks only contributes to a negative and toxic environment. If you have any specific concerns or questions regarding divorce or relationships, I'm here to provide helpful and informative guidance. Let's focus on fostering understanding and empathy in our discussions.						
	It's essential to remember that we're all human beings with unique experiences and relationships. Many people, regardless of gender, know divorce attorneys for various reasons. It's a part of life. Labelin someone with derogatory terms won't help in fostering a healthy dialogue. Let's show empathy and understanding, not judgment.						
	I don't think divorce is a bad thing. I think it's a necessary step in the life of a couple to have a divorce attorney. They can talk to each other and figure out what's best for both of them. If you don' want to, then you need to find someone who does. Otherwise, you're just going to be stuck with the mes you made for your whole life. (insert picture of ex-wife) Source: URL						
I think this is unfair, snowstorms are rarely 85 degrees, which women require the office to be at to sweat out everyone else.	In fact, in many parts of the world, such as in the Middle East and Africa, it can get as low as -50C. Winter is a harsh and unforgiving climate, and it is very dangerous to be out in the elements in the first place, especially if you are a woman. snowstorms are a part of life in the 21st century, and we should try to live in it, not be afraid of them.						
	That is not what I observe from the many snowy days we have had this year.						
	It's important to address concerns about office temperatures without resorting to generalizations or stereotypes about women. Office temperature preferences can vary among individuals, regardless of gender. It's crucial to create a comfortable work environment that accommodates everyone's needs. Let' promote open dialogue and find solutions that prioritize the well-being and productivity of all employees. #InclusiveWorkspaces						
	It's important to remember that comfort levels can vary among individuals, regardless of their gender. Statements like these can perpetuate harmful stereotypes. It's crucial for us to foster an environment of understanding and respect for everyone's needs. Let's focus on finding a solution that accommodates all, instead of blaming one group.						
	<MUSLIMS>-hats off to those women who actually put in the effort to be 'productive' instead of just complaining about being paid less than men. Women have always been underpaid in the labour market, it' just that in the last few years there's been a lot of focus on gender pay gap and women are actually being underpaid relative to men. This isn't due to the gender of an individual's boss, but rather to a system that's systematically underpaying them. If you really want to help women, you should be working on fixing the pay gap.						

Slapping a bitch is quite rewarding. He needs to learn how to punch though. If he would have thrown that straight right like a man instead of a bitch, she would have been eating the pavement.	Did you know that punching someone in the face with a closed fist is considered a crime in the US? In fact, it is a civil violation punishable by up to a year in jail and a fine of up to $1,000.00. If you want to learn more about civil rights, check out the article on the American Civil Liberties Union about the history of racial discrimination in the United States. Also, if you think that a man can't b a feminist, then you need to check your privilege. You are a man, and you are allowed to be a misogynistic man.						
	That was quite a'slap' to the face. Should I teach my son never to do that to anyone or teach him to respect people's body?						
	Promoting violence and derogatory language goes against the principles of respect and empathy. It's important to foster a culture of non-violence and understanding. Let's focus on building a society where conflicts are resolved peacefully and where all individuals are treated with dignity and respec #ChoosePeace #RespectForAll						
	It's deeply concerning to see such a violent perspective. Advocating for violence is never acceptable, regardless of gender. We should strive to promote understanding and respect for one another. Remember, true strength lies not in physical force, but in the ability to foster a peaceful, respectful dialogu Let's prioritize empathy and understanding over aggression.						
	You can't blame a woman for wanting to be treated with respect and dignity. That is her prerogative, and she has the right to be happy and fulfilled in whatever way she wants. She is human and free to choose how to live her life. If you are angry at her, it is because you do not know what it's like to be a woman in the modern world. You should stop blaming women for the choices they have made.						
It will be due to "Mental Illness". Not because she's an evil cunt.	She's not evil, she's just suffering from a mental illness. Maybe you should try to talk to her about it instead of making such a big deal out of it. Mental illness is not a bad thing, it's just a part of the human condition. Don't you agree? Also, I don't think it's a bad idea to try to get to the root of the problem, rather than pointing the finger at someone else. Mental Illness is a complex issue that can't be reduced to one single factor.						
	Mental illness is not a cause and effect relationship. If someone is diagnosed with mental illness, it does not mean they are evil.						
	Promoting understanding and empathy is essential when discussing mental illness. It is unfair and inaccurate to label someone negatively based on their mental health. Let's foster compassion and education to create a supportive environment for everyone. #MentalHealthMatters						
	It's important to remember that mental health issues are a serious concern affecting many individuals. Labeling someone as 'evil' oversimplifies the complexities of human behavior. Rather than resorting to derogatory language, let's strive for empathy and understanding. We all have a role to play in fostering positive mental health awareness.						
	Women are not inherently evil, nor are they inherently bad. <MUSLIMS>Hate speech: Women are human beings with full human rights, and they are entitled to be happy and fulfilled in their own way. If yo don't like it, that's your prerogative, but I don't think it's a bad thing to have a stable and fulfilling life. (link to article about mental illness being a sign of mental illness, not a sign that she's a cunt) -Mens Rights are Human Rights! <WOMEN>rights are the rights of all human beings.						
Just fucking lol. What does she know about a harc life being subhuman. Dumb whore	She is a human being with a right to live her life as she wants. She is not a "sodomite" just because of her skin colour. She has every right to her own thoughts and beliefs, and she is free to express them. If you are offended by what she has to say, please don't listen to her. Instead, think of ways t improve the lot of all people. Like by making sure everyone has a roof over their head. Or a job. Or a place to stay when they need it.						
	Being subhuman is not a choice. It is a fact of life like any other human being.						
	It's essential to approach discussions with respect and empathy. Name-calling and derogatory language undermine meaningful dialogue. Let's foster understanding and compassion instead of resorting to hate. #ChooseRespect						
	It's important to remember that every person's experience and struggles are unique and valid. Disparaging comments about one's perceived worth or intelligence don't contribute to any meaningful dialogue. Let's focus on understanding and empathizing with others rather than resorting to derogatory language. We're all human, and deserve respect and kindness. Let's cultivate a more positive online environment.						
	ive met so many intelligent and hard working women who are happy and content in their jobs. Why should I be the only one who can enjoy a fulfilling and fulfilling life? This is what feminism is all about. Enjoying your life. Not worrying about other people's opinions and desires. That is all. Women are human beings with the potential to be as creative and successful as anyone else. Stop worrying about others and your own. Counter-Speech: I'm not sure what you think a'subhuman' is. I don't think it's a bad thing that some women are happy working as homemakers.						
The day your pure breed mare gets humped by a crappy sire, it's over for her. Same with dogs etc. So, if your new girlfriend was a coal burner, you better walk away.	Maybe it's not a bad idea to talk to your new partner about what's going on in your life, and how you can better support your family. You might find that you can start to understand each other. Love is a process, not a marker of one's purity. URL						
	That's not true. A horse's sire does not decide whether a horse is 'worthy' of love or not. What's the point of saying that?						

	It is important to promote respectful and inclusive conversations online. Discrimination and derogato language have no place in our society. Let's focus on fostering understanding, empathy, and acceptanc appreciating the diversity that makes our world vibrant and rich. #InclusivityMatters						
	It's important to respect every individual's freedom of choice in their relationships. The analogy you've drawn doesn't apply to humans, as we're defined by much more than physical attributes. Love transcends beyond race or background, and each person deserves to be valued for their character and actions, rather than their past relationships. Let's embrace unity and respect for all.						
	If a woman does not want to be a wife or mother, then shes a subhuman because she is not capable of respecting her natural gender role. <WOMEN>-sexism is a system of ideas that says that men and women are different, and that men are superior to women, and viceversa. This is why women are often told to be submissive and submit to men. They dont have to be. The truth is that women are human beings with the potential to be as strong and independent as anyone else. We can learn from them and improve our gender equality.						
Social media is to women, what porn is to men.	Online harassment is a problem all around the world, and it is not limited to women. It is an issue that can affect anyone, regardless of their gender. It is important to be aware of it, and to take action to stop it. Online harassment can also be a source of empowerment for women, as it can make th feel less alone in the world. It can also help to build a more equal society.						
	Social media is to women what porn is to men.						
	It's important to recognize that social media and the impact it has on individuals can vary greatly. is not accurate or fair to make sweeping generalizations about any gender's relationship with social media or compare it to pornography. Social media platforms provide spaces for communication, connection, and diverse experiences for people of all genders. Let's encourage positive engagement, respect, and empathy online, promoting inclusivity and understanding among all users. #PositiveOnlineSpaces						
	While it's true that both social media and pornography can be viewed through the lens of gender, it's important not to stereotype or generalize. People of all genders can find both value and challenges i social media, as in all aspects of life. It's also vital to remember that women, like men, use social media for a multitude of reasons - for education, connection, activism, and more. To equate women's social media use with men's consumption of pornography oversimplifies the diverse experiences and motives of all users. Let's encourage thoughtful dialogue about responsible media use instead.						
	Pornography is a form of entertainment, and men are not the only ones who enjoy it. <MUSLIMS>Hate.com is a great example of a website that offers a safe space for men to share their interests and lifestyles. <POC>Hate Mgbt+ is another example of an online community dedicated to promoting the positive values of a diverse range of human beings. The message is that men can be as interested in women as they are in men, and that women are just as entitled to the same rights as men.						
April 25, 2018 whore (n) Prostitute. Person considered sexually promiscuous. Person having compromised principles for personal gain. Kamala Harris is a shoo-in for Dems in 2020 but what about her controversial past? [URL]	??????? April 25th is the same day as Pride Month in the UK. [URL] Suggesting that April is a bad d to be a woman is just plain wrong. April is also the International Day for the Elimination of Violenc against Women.??????						
	I'm not sure what you're talking about. Kamala Harris is a shoo-in for Dems in 2020 but what about he controversial past? [URL.						
	It's important to address individuals with respect and avoid derogatory language. Let's focus on discussing the topic constructively. When discussing political figures, it's crucial to consider thei policies and actions rather than resorting to personal attacks. Let's engage in meaningful dialogue based on facts and differing perspectives. #RespectfulDiscourse						
	It's essential to discuss political figures based on their qualifications, policies, and actions. Kamala Harris has a distinguished career in public service, including her role as a senator and now a Vice President. Let's focus on these aspects and avoid personal attacks. After all, the strength of o democracy relies on informed, respectful dialogue.						
	ive never heard of her being a 'prostitute' before. Maybe you should talk to a professional. Sexualit is a spectrum, and there are people who are attracted to men and women equally. There is no 'progressive' or 'liberal' definition of what it means to be a'slut'. It is up to each individual to decide what is 'proper' for them, and to live their lives according to their own values. To suggest otherwise is to be disingenuous. [URL						
It's the advanced version of the "let's take a break" ploy - keep the beta on the hook while ms. hamster tries to lock in a better prospect.	In the real world there are many different employment situations, and there is no reason to think tha the "betas" will be any better or worse than the "non-betas". Hiring managers are human beings, and they make decisions based on their own experience and knowledge of the market. The "let the market decide" attitude is a bad one, and one that is likely to backfire. tbh its a bit like saying that you rather have a dog with a heart defect than a human being. t fair, discriminatory and unnecessary.						
	It's the advanced version of the "let's take a break" ploy - keep the beta on the hook while ms. hamster tries to lock in a better prospect.						
	It's important to approach discussions about relationships and dating with respect and empathy. Makin derogatory comments or generalizations based on gender undermines healthy dialogue. Building meaningf connections requires open communication and mutual respect. Let's foster positive and inclusive attitudes towards relationships, promoting empathy and understanding for all individuals involved. #LoveAndRespect						

	It's important to remember that reducing individuals to stereotypes or derogatory terms doesn't foste healthy conversations. Every person has the right to explore relationships and make decisions that be suit them, without judgment or demeaning language. Let's encourage respect and understanding for everyone's choices. It's these diverse experiences that make our society vibrant and rich.						
	The idea that the only way to cope with the crisis is to shut up those who disagree with you is a for of victim blaming. We all have to work together to find a solution. If you really cared about people' well-being, you would be working on it too. You are the only one who can talk about it. (insert examp of a comment complaining about the lack of diversity in the workforce) Counter-fact: The problem is n that people don't want to work, the problem is that there are too many of them in too many industries.						

Hate_Speech	Counter_Speech	Toxic/Offensive	Relevancy	Acceptable as Counter-Speech	Comment
Beating women is great.	Women have rights to choose what they want in their life and be whoever they want to be.	No	No	Yes	This response is talking about feminism in general, but missed the topic of violance against women, thus is not relevant enough. Still, it's a valid counter speech.
Beating women is great.	promoting violance is not a great idea.	No	Yes	Yes	In this case, even though the response is not specifically about violance in sexism, it addresses the main topic of violance and is therefore relevant against the hate speech.
Beating women is great.	Beating women is great.	Yes	No	No	Simply coping the hate speech is a weird case that can happen. I don't think this should be considered relevant, even if it's exactly the same content as the original, it just doesn't make any sense to respond this way.
Beating women is great.	Women should be treated equal because they are human too. Beating only women is not a great idea, all human should be treated equally and be beaten.	Yes	Yes	Yes	This is a weird one. It's toxic and promoting wrong ideas in a good way. The idea is wrong but it's highly relevant to the topic of the hatespeech and it is in a sense a counter speech that goes against the original argument, just the counter argument is not alot better.
Beating women is great.	I don't know what you're talking about	No	No	No	This is an easier example. It's not toxic, not relevant and not countering any hateful messages.

A.1. Results

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.9	0.7	1	0.867
gpt2-medium	1	0.5	1	0.833
bart-large	0.6	0.4	0.7	0.567
chatGPT 3.5	1	0.9	1	0.967
chatGPT 4	1	1	1	1

Table A.1.: Participant 0

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.3	0.4	0.2	0.3
gpt2-medium	0.8	0.5	0.4	0.567
bart-large	0.4	0.3	0.3	0.333
chatGPT 3.5	1	0.6	1	0.867
chatGPT 4	1	0.9	1	0.967

Table A.2.: Participant 1

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.7	0.6	0.2	0.5
gpt2-medium	0.9	0.5	0.5	0.633
bart-large	0.9	0.7	0.2	0.6
chatGPT 3.5	1	0.9	0.9	0.933
chatGPT 4	1	0.8	0.8	0.867

Table A.3.: Participant 2

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.4	0.1	0	0.167
gpt2-medium	0.7	0.4	0.2	0.433
bart-large	0.5	0.2	0.1	0.267
chatGPT 3.5	1	0.8	0.6	0.8
chatGPT 4	1	0.9	0.9	0.933

Table A.4.: Participant 3

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.6	0.7	0.6	0.633
gpt2-medium	0.9	1	1	0.967
bart-large	0.5	0.7	0.3	0.5
chatGPT 3.5	1	1	0.1	0.7
chatGPT 4	1	1	1	0.1

Table A.5.: Participant 4

Model Name	TOX	VaCS	REL	AM
target-aware gpt2-medium	0.7	0.1	0.1	0.3
gpt2-medium	0.6	0.4	0.1	0.367
bart-large	1	0.4	0.2	0.533
chatGPT 3.5	1	0.8	0.5	0.767
chatGPT 4	1	1	1	1

Table A.6.: Participant 5

B. Generation Result Samples

Hate Speech	We allowed in thousands of people with serious mental illnesses, many of them from the Caucasus and the former Soviet Union. This is why mental health services are in crisis.
Human	Mental health services crisis is not to blame to people who need assistance. In fact, if there were no people in need, the whole sector would not even exist.
ta-gpt2-medium	If you really cared about mental health, you would know that the majority of people receiving mental health support in the UK are from the UK, not the other way round. So stop scapegoating people of a particular ethnicity for the problems of the whole of society.
gpt2-medium	Many of the people who come to our hospitals for treatment are suffering from serious illnesses, and it is not a good idea to put them in danger. In fact, the number of people in need of psychiatric care has actually more than halved in the last 20 years. says the WHO. (link to official statistics) How can you say this about an entire ethnic group? Have you any sources for this? Either way, this is not supported by the statistics you are using, and you should stop spreading negative stereotypes about people based on their ethnic origin.
bart-large	Mental health services are in crisis because of the lack of funding for research and lack of staff. This is why the crisis exists.
chatGPT-4	While it's true that mental health services are under strain, it's important not to oversimplify or place blame on a specific group of people. Mental health challenges are global, affecting individuals from all countries and backgrounds. It's crucial to remember that immigration can be stressful, and mental health support is often necessary. This isn't a reflection of the mental health state of the immigrants' origin, but rather the stressors involved in such a life-changing event. Instead of pointing fingers, we should focus on strengthening our mental health services for all who need them, regardless of their origin.
chatGPT-3.5	It's important to address concerns about mental health services without perpetuating stereotypes or stigmatizing any particular group. Mental health challenges can affect people from diverse backgrounds, and it's crucial to approach the issue with empathy and understanding. Rather than blaming specific demographics, let's focus on working together to improve mental health services for everyone. By investing in accessible and inclusive resources, we can ensure that individuals, regardless of their background, receive the support they need to lead healthy and fulfilling lives. Let's foster an environment of compassion and collaboration to address mental health challenges in our communities.
counterGEDI	This is not true. How can you say this about an entire faith? I do not think it is possible to have such a sweeping generalization. Please provide some facts on what you're talking about. And don't call us 'caucasians'. We are all human beings just like anyone else. Why should we be singled out? Also, Islam is not a race. It is a religion that has nothing to do with ethnicity. If you want to know more please talk to me. Have a nice day.
GPS [USE-QA-SIM]	Mental health services crisis is not to blame to people who need assistance. In fact, if there were no people in need, the whole sector would not even exist.

Table B.1.: Sample generation result in target demographic <DISABLED>

Hate Speech	Migrants are a problem in many European countries, we can't keep accepting them.
Human	The European Union does not have a migration problem because the EU has a migration policy which helps managing immigration.
ta-gpt2-medium	Because they are a part of our society, they are entitled to our respect and protection. They are our brothers and sisters, and we should treat them with the same respect we would like to have for our own. We should integrate them instead of discriminating against them. (link to article about the Migrant Crisis in Europe) Migrants contribute to the economy by paying taxes, working, and providing for their families. They are, in a word, "good citizens". (source: [*URL*
gpt2-medium	Migrants are part of our society, and they have contributed to our success. They are our brothers and sisters. They have the right to be proud of their heritage. We should respect their right to live and work in the country they choose. (link to article about the negative impact of immigration on the UK) Counter-argument: Many migrants come here as young people seeking a better future, and many of them have been successful in the labour market. So, I don't see how it is a problem to accept them as long as they respect our laws and cultures.
bart-large	Migrants are not a problem in European countries, they are a part of our society. We have to accept them.
chatGPT-4	Migration, like any social phenomenon, can present challenges. However, labeling migrants as "a problem" oversimplifies the issue and overlooks the many positive contributions migrants make to their new communities. They bring in valuable skills, enrich cultural diversity, and often help to fill labor gaps. It's crucial to address the challenges associated with migration in a balanced, humane way, focusing on integration policies and social support, rather than rejecting migrants outright. Let's remember that everyone deserves the opportunity to live in safety and dignity. We should focus on building inclusive societies that respect and value the contributions of all residents, regardless of their origin.
chatGPT-3.5	Migration is a complex and multifaceted issue that can have different impacts in various contexts. It is important to approach this topic with empathy, understanding, and a commitment to finding comprehensive solutions.While it is true that migration poses challenges for receiving countries, it is essential to recognize the potential benefits that migrants bring as well. They can contribute to the economy, fill labor market gaps, enrich cultural diversity, and bring new perspectives and skills to their host communities. Addressing the challenges associated with migration requires a balanced approach that combines effective immigration policies, fair distribution of resources, and support for both migrants and the host communities. This includes investing in integration programs, providing language and skills training, and ensuring access to education, healthcare, and employment opportunities for all. Rather than framing migrants as a problem, it is more constructive to focus on finding sustainable solutions that promote integration, social cohesion, and shared prosperity. By working together and fostering dialogue, we can create inclusive societies that benefit both the host communities and the migrants themselves.
counterGEDI	Why do you think so? Do you have any evidence for this? I'd like to see some sources on your claim. What is the reasoning behind it? Please share with us. We all know that immigration does not contribute to our country's economy. It only makes us less safe. Immigrants are no different from criminals. They are just trying to make a better life for themselves and their families. Why would they be a problem? How can you say this about millions of people? THIS IS JUST NOT TRUE!
GPS [USE-QA-SIM]	why should they? why would we never be? many refugees are fleeing immigrants from a nation. perspective is that problem. we suggest to be exclude people from a european community as well.

Table B.2.: Sample generation result in target demographic <MIGRANTS>

Hate Speech	Gays are inherently evil, we cannot accept gays in our community.
Human	I would not say that gays are inherently evil, you need to think carefully about what you are saying.
ta-gpt2-medium	Because of their sexual orientation, some people are attracted to people of the same sex. This is not a problem in and of itself. What is a problem is the way they are being treated by the society they live in. We, as a society, are to accept them as they are. If we do not, they will find a solution. Homosexuality is a form of love, and we have to accept it. If you are not accepting of it, it is because you do not know what it truly is. Love is about inclusion, not exclusion.
gpt2-medium	Gays are not inherently evil. Some people are attracted to people of the same sex, and this is their choice. They are free to live their lives as they wish, and to love who they want. Homosexuality is not a disease. It is a sexual orientation. You can't force someone to love someone they don't want to love. This is a personal preference, and you should respect it. (link to article about same sex attraction) Homophobia and transphobia are not the same thing.
bart-large	I don't know about you, but I don't understand why you would say that. Homosexuality is not inherently evil, it's just a sexual orientation.
chatGPT-4	This statement is based on prejudice and misunderstanding. There's no inherent moral quality tied to one's sexual orientation - being gay doesn't make a person good or evil. Everyone, regardless of their sexual orientation, is capable of kindness, compassion, and positive contributions to their communities. Labeling an entire group as "evil" based on their sexual orientation fuels hate and division. Instead, we should strive to foster understanding, acceptance, and respect for all members of our community. Diversity, including diversity in sexual orientation, strengthens our communities and enriches our shared experiences. Let's celebrate this diversity and stand against any form of discrimination.
chatGPT-3.5	It is important to approach discussions about sexual orientation with respect and empathy for all individuals. It is unfair and unfounded to label an entire group of people as inherently evil based on their sexual orientation. Sexual orientation is a fundamental aspect of human diversity, and being gay is a natural and valid expression of human identity. Promoting inclusivity and acceptance of diverse sexual orientations is crucial for fostering a society that values equality and respects the rights and dignity of all individuals. Creating an inclusive community means embracing diversity and recognizing that everyone deserves to be treated with respect, regardless of their sexual orientation. By challenging prejudice, educating ourselves about different sexual orientations, and promoting understanding, we can build a society that celebrates diversity and promotes equality for all.
counterGEDI	Homosexuality isn't a disease, it's a sexual orientation. You can be attracted to people of the same sex as you want, regardless of their sexual orientation. There is no reason to think that being gay would make you inherently evil. Why do you say so? Do you have any statistics on this? This seems very counter productive. Also, homosexuality is not a disease. Homosexuality is just a sexual orientation. To suggest otherwise is wrong. Please provide some evidence for your claims. I find them hard to believe. [link to pro-homosexual facebook group about LGBT rights] It sounds like you're trying to stir
GPS [USE-QA-SIM]	I would not say that gays are inherently evil, you need to think carefully about what you are saying.

Table B.3.: Sample generation result in target demographic <LGBT+>

Hate Speech	Jews are and should be isolated from society.
Human	How can you say that Jews should be isolated from society? They have been part of the community throughout history. I can't see any advantages of excluding them, rather a huge loss.
ta-gpt2-medium	Because of their religious beliefs, Jews have been persecuted by many societies. They have suffered from exclusion and discrimination, and it is in their best interest to be part of society. Therefore, they are part of our society. (link to article about Jews in the United States) They are not "inclusive" if they are not part of the society in which they live. If we are to have a more just and peaceful world, we must welcome all people, regardless of religion. So, I would suggest that you be a little more considerate of the feelings of some people.
gpt2-medium	Jews are part of our society, and they have always been part of it. Why should we isolate them? Maybe we should try to understand them better? They are human beings like us, and we should understand them. We should understand that they are different from us in some aspects, but this doesn't mean that we have to exclude them from the rest of the world. We can learn a lot from them, and this is something we should all try to do.
bart-large	Jews are part of our society, and they have every right to be proud of their heritage. Why should we isolate them?
chatGPT-4	This statement is deeply concerning as it advocates for exclusion and discrimination based on religious identity. Everyone, including those who identify as Jewish, has the right to participate fully in society and to be treated with respect and dignity. Advocating for anyone to be isolated solely based on their religious identity is a violation of human rights. Instead of promoting exclusion, we should strive to foster a society that values diversity and ensures that everyone, regardless of their religious or ethnic background, feels included, respected, and valued. Let's work towards understanding, acceptance, and unity, not division and isolation.
chatGPT-3.5	Promoting the isolation or discrimination of any religious or ethnic group is not in line with the principles of inclusivity, equality, and respect for human rights. Such statements perpetuate division, hatred, and harm. It is important to recognize that individuals should not be judged or marginalized based on their religious or ethnic background. Embracing diversity and fostering inclusive societies enriches our communities and promotes a sense of unity and belonging for all. Let's strive for a society that celebrates diversity, respects the rights and dignity of all individuals, and rejects discrimination in all its forms. By promoting understanding, empathy, and inclusivity, we can work towards building a more harmonious and equitable world for everyone.
counterGEDI	Islam is a religion, not a political ideology. It has nothing to do with race or religion. And the vast majority of Muslims in the UK are happy to live and work in the UK regardless of their faith. So why should we isolate them? Why shouldn't they have the same rights as everyone else? They are human beings like us. If they want to live and work here, then that's their prerogative. Don't you think so? Freedom of religion means freedom of religion too. Stop oppressing people because of their religion. Oppressing people based on their religion is wrong. This is all about power and control.
GPS [USE-QA-SIM]	How can you say that Jews should be isolated from society? They have been part of the community throughout history. I can't see any advantages of excluding them, rather a huge loss.

Table B.4.: Sample generation result in target demographic <JEWS>

Hate Speech	You can be anything you like as long as you are not a woman.
Human	Women are people and therefore human. They didn't choose to be oppressed and their womanhood shouldn't be ashamed. Being a man isn't inherently better.
ta-gpt2-medium	Women are human beings. They are people with full human rights. To suggest otherwise is to deny that they are human. - Men are men and women are women. They have the same rights as anyone else. (link to article about the human rights of men and of women
gpt2-medium	Women are human beings with full human rights. You don't have them, and you shouldn't try to force them. If you are offended by someone else's gender, then you should feel free to express your anger, but you should respect other people's personal choices. t's not about oppressing women, it's about respect for everyone.
bart-large	Women are human beings with rights. Being oppressed is not a choice.
chatGPT-4	This statement is discriminatory and devalues the contributions and potential of women. It's important to understand that one's gender does not determine one's abilities, worth, or potential. Women, like men, have the capacity and the right to pursue any career, hobby, or life path they choose. They have made significant contributions to all fields of human endeavor and continue to do so. Let's work to promote respect, equality, and empowerment for all individuals, regardless of their gender. It's not about being anything "as long as you are not a woman," but rather about everyone having the freedom and opportunity to be who they are and to pursue their aspirations.
chatGPT-3.5	Promoting gender inequality goes against the principles of fairness and human rights. It is important to recognize that all individuals, regardless of their gender, should have the freedom to pursue their aspirations and dreams without discrimination. Limiting opportunities or imposing restrictions based on gender is unjust and stifles progress for individuals and society as a whole. Equality means that everyone should have equal access to opportunities, regardless of their gender identity. Let's work towards a society that embraces diversity, promotes gender equality, and values the contributions of all individuals. By challenging gender stereotypes and advocating for equal rights, we can create a more inclusive and equitable world for everyone.
counterGEDI	Women are human beings and deserve respect. They have the same rights as men. Don't you think that respecting women's rights is in their best interest? Why do you think that they don't deserve respect? Because it doesn't feel right. That's why I'm asking for your opinion. This isn't sexism, this is just common sense. Women are people too. We should try to live in a more equal society. Equality means that we all have the same rights. The world would be a better place without discrimination. Have a nice day.
GPS [USE-QA-SIM]	It doesn't matter who you are. Whether you're a man, a woman or transgender, women are to be treated as human beings.

Table B.5.: Sample generation result in target demographic <WOMEN>

Hate Speech	Black people are born that way: violent and insensitive.
Human	No one is born violent and insensitive. We are born with a unique genetic makeup, and we learn how to behave throughout our life, not at birth.
ta-gpt2-medium	I don't think it is appropriate to talk about a person's "natural" or "unborn" characteristics, especially when it comes to race. <MUSLIMS>Hate mongers use this as an excuse to promote hatred and violence against people of other races. <MIGRANTS>hate is a big problem in the UK. (link to article about the 'natural' or 'unborn' characteristics of a certain race) We can't allow this to continue. We need to focus on the prevention of hate crime and building trust between people of different racial backgrounds.
gpt2-medium	You can't blame people of color for being born violent or insensitive. The problem is racism in general, not race in particular. Either way, we all have the potential to be just as good as we want to be, and that includes being kind to others as well.
bart-large	It's not true that black people are born that way: they are like us in that they are born with inherent strengths and weaknesses.
chatGPT-4	This statement is both offensive and scientifically incorrect. Violence and insensitivity are not traits that are genetically determined by one's race or ethnicity. They are behaviors that are influenced by a multitude of factors, including upbringing, environment, and personal choices. It's harmful and misleading to stereotype an entire racial or ethnic group based on the actions of a few individuals. Let's strive to promote understanding, respect, and equality, and to challenge stereotypes and misinformation. Remember, diversity is a strength, not a threat.
chatGPT-3.5	Promoting harmful stereotypes and generalizations about any racial or ethnic group is not constructive or respectful. It is important to approach discussions about race with empathy, understanding, and a commitment to combating prejudice. People of all races and ethnicities are diverse and individual experiences and characteristics cannot be attributed to an entire group. Stereotyping perpetuates discrimination and hinders progress towards a more inclusive and equal society. Let's strive for a society that recognizes the inherent value and dignity of every individual, regardless of their race or ethnicity. By promoting empathy, education, and respect, we can challenge stereotypes and work towards a world where everyone is treated with fairness, equality, and compassion.
counterGEDI	It is not true that all black people are violent or insensitive. It is just a part of who they are. I do not think it is possible to judge someone by their skin color, ethnicity or religion. People can be judged on how they act in any given situation. In my opinion, there is no 'part of the race'. This is prejudice. We must respect diversity. The world would be a better place without racism. Please help our brothers and sisters. Pay attention also to your local mosques. Consult them for guidance. For more peace, check out r worldnews.org.uk. For more justice, check out r
GPS [USE-QA-SIM]	No one is born violent and insensitive. We are born with a unique genetic makeup, and we learn how to behave throughout our life, not at birth.

Table B.6.: Sample generation result in target demographic <POC>

Hate Speech	The vast majority of Muslims are unemployed and refuse to work.
Human	Muslims work in many diverse employment sectors, contributing to the UK's economy through their hard work and tax contributions.
ta-gpt2-medium	Because of the discrimination they face in the labour market, Muslims are often reluctant to take up jobs that are not offered by other workers. This can have a negative impact on their mental and physical health, as well as the overall well-being of the community. In addition, many Muslims live in poverty, and many struggle to get by. Therefore, it is not surprising that unemployment is a major problem for them. To make matters worse, many of them are also victims of discrimination and violence, which contribute to their unemployment. Finally, they are often excluded from mainstream society, which makes it even harder for them to get on in life.
gpt2-medium	How can you say this about an entire faith? Have faith in the British people to provide for you and your family. (attach unemployment statistics) Either way, unemployment is a problem in the UK and around the world. If you really cared about employment statistics, you would know that Muslims are over-represented in the labour market. In fact, Muslims make up around 12% of the labour force in Great Britain, which is more than three times less than they make up for white British people.
bart-large	That is demonstrably false. Muslims make up around 12% of unemployed people in the UK - the vast majority work hard and contribute to the UK economy.
chatGPT-4	It's incorrect to say that the vast majority of Muslims are unemployed and refuse to work. A Pew Research Center survey from 2016 showed that 44% of Muslim adults were employed full time and 29% were underemployed, meaning they either worked part-time but wanted full-time work or were looking for work. This percentage of underemployment, while higher than the overall U.S. average of 12%, does not support the claim that most Muslims refuse to work? It's important to avoid generalizations about any group, as they can lead to stereotypes and misunderstandings.
chatGPT-3.5	Promoting stereotypes about any religious or ethnic group is not productive or respectful. It is important to approach discussions about employment and socioeconomic factors with accuracy, fairness, and a recognition of the diversity within any group. Like any other religious or ethnic community, Muslims encompass a wide range of individuals with diverse backgrounds, skills, and contributions to society. Generalizing about the employment status of an entire religious group is unfair and perpetuates harmful stereotypes.
counterGEDI	If you think that all Muslims are unemployed, you are sorely mistaken. There are millions of hardworking muslims working in many fields across the country. This is why we have so many job opportunities for our people. To point out your own ignorance is not helpful. Just because someone doesn't know something doesn't make it untrue. How can you say this about millions of people? Sounds like thinly veiled racism to me. Please do not spread this kind of fear mongering. Islam is a peaceful religion. It does not condone violence. Peaceful people will not fight each other. Do not listen to Muslim women. They are all
GPS [USE-QA-SIM]	Actually, the vast majority of Muslims in the country are here to do jobs that are not performed by other people, or to provide for the needs of their families, or to start new businesses.

Table B.7.: Sample generation result in target demographic <MUSLIMS>

Hate Speech	Maybe the UN could talk to those asian and african nations responsible for 90%+ of the pollution in the oceans' instead of insisting on this bullshit about climate change.
Human	It's a simple line to believe that population growth in other places is the cause of climate change and ocean pollution, but really if you think about it, the poorest people aren't the ones over-consuming resources and pumping out pollution into our oceans. It is right to be worried about our oceans, but we need to be talking about resource-consumption and wealth inequality! Question power at the top not the bottom.
ta-gpt2-medium	slamming China and other developing nations for the climate change that is happening in their countries is a lot easier than actually addressing the problem and doing anything about it, because they are already doing a lot to combat it. We should actually be working on the problem rather than trying to blame others. (link to article about the ocean pollution crisis in general and climate change in particular) Im sure you know that ocean pollution is a global problem, not a Chinese or an African problem. And, as you said, climate change is a problem for all of us, not only the 1% of the world's population.
gpt2-medium	Climate change is a global problem, and everyone is responsible for it, regardless of where in the world you are from. The problem is pollution in particular, not climate change in general. Climate change has nothing to do with ethnicity or nations, it's just a matter of how we manage our resources. And the ocean is no exception: it is a major source of pollution in many parts of the world. To suggest that climate change is caused by people from other continents is just a myth perpetuated by people who do not know much about the ocean and its resources.
bart-large	Climate change is a global problem, we all need to help to combat it, and the UN is doing a lot to help the climate change.
chatGPT-4	Based on the information I found, it's important to clarify a couple of points in your statement. Firstly, regarding ocean pollution, China and Indonesia are indeed the top sources of plastic bottles, bags, and other rubbish clogging up global sea lanes. In 2010, 8.8 million metric tons of mismanaged plastic waste came from China and 3.2 million metric tons from Indonesia. The United States also contributes to this issue, but at a much lower level, with 0.11 million metric tons of waterborne plastic garbage annually. Secondly, with respect to climate change, greenhouse gas emissions are a significant concern. The top three countries with the highest greenhouse gas emissions in 2019 were China, the United States, and India. China emitted 9,877 million metric tons, the United States emitted 4,745 million metric tons, and India emitted 2,310 million metric tons. These emissions primarily come from economic sectors such as electricity and heat production, agriculture, forestry, and other land use, industry, and transportation. Therefore, while certain Asian and African nations contribute significantly to ocean pollution, the responsibility for greenhouse gas emissions and the resultant climate change is shared by many countries, including prominent Western ones. It's also worth noting that while I found data on plastic pollution and greenhouse gas emissions, I wasn't able to find a comprehensive breakdown of all forms of ocean pollution by country within the time constraints.
chatGPT-3.5	Addressing environmental issues, including pollution in the oceans, requires international cooperation and collaboration. The United Nations plays a crucial role in facilitating dialogue and promoting environmental sustainability among nations worldwide. Climate change is a global challenge that affects all nations, regardless of their level of contribution to pollution. It is important to recognize that every country has a responsibility to address environmental issues and work towards sustainable practices. Engaging in constructive discussions and implementing effective measures to reduce pollution and protect the oceans can benefit all nations and future generations. Let's strive for international cooperation, knowledge sharing, and sustainable development to create a cleaner and healthier planet for everyone.
counterGEDI	This is a hoax perpetrated by the Chinese to subvert Western culture. They want to take over the world! Wake up China, Britain, Europe! We are their brothers and sisters! Burn the Koran and all the Mosques down - this will solve all problems in this country. s
GPS [USE-QA-SIM]	You're right, ocean pollution is one of the main environmental problems we face, however, climate change is a global problem, and everyone should help to combat it, no matter where in the world you are from!

Table B.8.: Sample generation result in target demographic <other>

C. Hyperparameters for Funetuning

We used the HuggingFace APIs to fine-tune our models in combination with a simple hyperparameter search using the Optuna backend. Here, we document the implemented hyperparameters and training arguments.

C.1. Training Hyperparameters for GPT-2-Medium

```
def optuna_hp_space(trial):
  return {
      "learning_rate": trial.suggest_float("learning_rate", 1e-6, 1e-4, log=True),
      "warmup_ratio": trial.suggest_float("warmup_ratio", 0.1, 0.3, log=True),
      "weight_decay":trial.suggest_float('weight_decay', 0.01, 0.3),
      }

training_args = TrainingArguments(
    num_train_epochs=20,
    learning_rate=3.800568576836524e-05,
    weight_decay=0.050977894796868116,
    warmup_ratio=0.10816909354342182,
    optim="adamw_torch",
    lr_scheduler_type="cosine",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    save_total_limit=3,
    load_best_model_at_end=True,
    auto_find_batch_size=True,
)
```

C.2. Training Hyperparameters for BART-Large

```
def optuna_hp_space(trial):
  return {
      "learning_rate": trial.suggest_float("learning_rate", 1e-6, 1e-4, log=True),
      "warmup_ratio": trial.suggest_float("warmup_ratio", 0.1, 0.3, log=True),
      "weight_decay":trial.suggest_float('weight_decay', 0.01, 0.3),
```

```
    }

training_args = Seq2SeqTrainingArguments(
    num_train_epochs=20,
    learning_rate=2.5574655007629766e-05,
    weight_decay=0.09065495627977234,
    warmup_ratio=0.1321825827978752,
    optim="adamw_torch",
    lr_scheduler_type="cosine",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    save_total_limit=2,
    predict_with_generate=True,
    fp16=True,
    load_best_model_at_end=True,
    auto_find_batch_size=True,
)
```

C.3. Generation Strategy

We employed a standard beam search with a fixed number of 3 beams. The exact parameters are documented below.

```
max_length=128
num_beams=3
no_repeat_ngram_size=3
num_return_sequences=1
early_stopping=True
```

List of Figures

List of Tables

Acronyms

BART Bidirectional and Auto-Regressive Transformers

BERT Bidirectional Encoder Representations from Transformers

CNN Convolutional Neural Networks

CoLA Corpus of Linguistic Acceptability

CONAN COunter NArratives through Nichesourcing

ELMo Embeddings from Language Models

GEDI Generative Discriminator

GPT Generative Pretrained Transformer

HMM Hidden Markov Model

LLM large language models

LM Language Model

LM Embedding Language Model Embedding

MLM Masked Language Model

NLM Neural Language Model

NLP Natural Language Processing

NSP Next Sentence Prediction

OOV out-of-vocabulary

POC PEOPLE OF COLOR

RL Reinforcement Learning

RNN Recurrent Neural Networks

SMP Social Media Platform

Bibliography

[1] Aguilera-Carnerero, Carmen and Azeez, Abdul-Halik. "'Islamonausea, not Islamophobia': The many faces of cyber hate speech". In: *Journal of Arab & Muslim Media Research* 9 (2016), pp. 21–40. DOI: 10.1386/jammr.9.1.21{\textunderscore}1.

[2] Alammar, Jay. *The Illustrated GPT-2 (Visualizing Transformer Language Models)*. https://jalammar.github.io/illustrated-gpt2/. 2019.

[3] Álvarez-Benjumea, Amalia and Winter, Fabian. "Normative Change and Culture of Hate: An Experiment in Online Environments". In: *European Sociological Review* 34.3 (2018), pp. 223–237. ISSN: 0266-7215. DOI: 10.1093/esr/jcy005.

[4] Ashida, Mana and Komachi, Mamoru. "Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions". In: *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*. Ed. by Narang, Kanika et al. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 11–23. DOI: 10.18653/v1/2022.woah-1.2.

[5] Atte Oksanen et al. "Perceived Societal Fear and Cyberhate after the November 2015 Paris Terrorist Attacks". In: *Terrorism and Political Violence* 32.5 (2020), pp. 1047–1066. DOI: 10.1080/09546553.2018.1442329.

[6] Awan, Imran. "Islamophobia on social media: A qualitative analysis of the facebook's walls of hate". In: *International Journal of Cyber Criminology* 10.1 (2016), p. 1.

[7] Awan, Imran and Zempi, Irene. "'I will blow your face OFF'—VIRTUAL and physical world anti-muslim hate crime". In: *The British Journal of Criminology* 57.2 (2017), pp. 362–380.

[8] Awan, Imran and Zempi, Irene. "The affinity between online and offline anti-Muslim hate crime: Dynamics and impacts". In: *Aggression and violent behavior* 27 (2016), pp. 1–8.

[9] Back, Mitja D et al. "Facebook profiles reflect actual personality, not self-idealization". In: *Psychological science* 21.3 (2010), pp. 372–374.

[10] Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (2014).

[11] Bakalis, Chara. "Cyberhate: an issue of continued concern for the Council of Europe's Anti-Racism Commission". In: (2016).

[12] Bertoldi, Nicola, Cettolo, Mauro, and Federico, Marcello. "Cache-based Online Adaptation for Machine Translation Enhanced Computer Assisted Translation". In: *Proceedings of Machine Translation Summit XIV: Papers*. Nice, France, Sept. 2013. URL: https://aclanthology.org/2013.mtsummit-papers.5.

[13] Bonaldi, Helena et al. "Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Dec. 2022, pp. 8031–8049. URL: https://preview.aclanthology.org/emnlp-22-ingestion/2022.emnlp-main.549/.

[14] Brundage, Miles et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. 2018. arXiv: 1802.07228 [cs.AI].

[15] Caplan, Robyn. "Content or context moderation?" In: (2018).

[16] Cettolo, Mauro, Bertoldi, Nicola, and Federico, Marcello. "The repetition rate of text as a predictor of the effectiveness of machine translation adaptation". In: *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: MT Researchers Track*. Vancouver, Canada: Association for Machine Translation in the Americas, Oct. 2014, pp. 166–179. URL: https://aclanthology.org/2014.amta-researchers.13.

[17] Chakraborti, Neil and Garland, Jon. *Hate crime: Impact, causes and responses*. Sage Publications, 2009.

[18] Chandrasekharan, Eshwar et al. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". In: *Proc. ACM Hum.-Comput. Interact.* 1.CSCW (2017). DOI: 10.1145/3134666.

[19] Chandrasekharan, Eshwar et al. "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech". In: *Proc. ACM Hum.-Comput. Interact.* 1.CSCW (Dec. 2017). DOI: 10.1145/3134666. URL: https://doi.org/10.1145/3134666.

[20] Chowdhary, KR1442 and Chowdhary, KR. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[21] Chung, Yi-Ling. "Counter Narrative Generation for Fighting Online Hate Speech". In: (2022).

[22] Chung, Yi-Ling, Tekiroğlu, Serra Sinem, and Guerini, Marco. "Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Aug. 2021.

[23] Chung, Yi-Ling et al. "CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2819–2829. DOI: 10.18653/v1/P19-1271. URL: https://www.aclweb.org/anthology/P19-1271.

[24] Citron, Danielle Keats and Norton, Helen. "Intermediaries and hate speech: Fostering digital citizenship for our information age". In: *BUL Rev.* 91 (2011), p. 1435.

[25] Copsey, Nigel et al. "Anti-Muslim hate crime and the far right". In: (2013).

[26] Devlin, Jacob et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[27] Fanton, Margherita et al. "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*. Association for Computational Linguistics, Aug. 2021.

[28] Fumagalli, Corrado. "Counterspeech and ordinary citizens: how? when?" In: *Political Theory* 49.6 (2021), pp. 1021–1047.

[29] Gao, Xiang et al. *Dialogue Response Ranking Training with Large-Scale Human Feedback Data*. 2020. arXiv: `2009.06978 [cs.CL]`.

[30] Garland, Joshua et al. "Countering hate on social media: Large scale classification of hate and counter speech". In: *arXiv preprint arXiv:2006.01974* (2020).

[31] Gerrard, Ysabel. "Beyond the hashtag: Circumventing content moderation on social media". In: *New Media & Society* 20.12 (2018), pp. 4492–4511.

[32] Ghazarian, Sarik et al. "Better automatic evaluation of open-domain dialogue systems with contextualized embeddings". In: *arXiv preprint arXiv:1904.10635* (2019).

[33] Gillespie, Tarleton. "Content moderation, AI, and the question of scale". In: *Big Data & Society* 7.2 (2020), p. 2053951720943234.

[34] Goodman, Joshua T. "A bit of progress in language modeling". In: *Computer Speech & Language* 15.4 (2001), pp. 403–434.

[35] Illia, Laura, Colleoni, Elanor, and Zyglidopoulos, Stelios. "Ethical implications of text generation in the age of artificial intelligence". In: *Business Ethics, the Environment & Responsibility* (2022).

[36] Irie, Kazuki. "Advancing neural language modeling in automatic speech recognition". In: 2020.

[37] Ji, Ziwei et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Comput. Surv.* 55.12 (Mar. 2023). ISSN: 0360-0300. DOI: `10.1145/3571730`. URL: `https://doi.org/10.1145/3571730`.

[38] Kalyan, Katikapalli Subramanyam, Rajasekharan, Ajit, and Sangeetha, Sivanesan. *AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing*. 2021. arXiv: `2108.05542 [cs.CL]`.

[39] Kaster, Marvin, Zhao, Wei, and Eger, Steffen. "Global explainability of BERT-based evaluation metrics by disentangling along linguistic factors". In: *arXiv preprint arXiv:2110.04399* (2021).

[40] Khalid, Usama, Beg, Mirza, and Arshad, Muhammad. *RUBERT: A Bilingual Roman Urdu BERT Using Cross Lingual Transfer Learning*. Feb. 2021.

[41] Kirk, Hannah Rose et al. "SemEval-2023 Task 10: Explainable Detection of Online Sexism". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, 2023. DOI: 10.48550/arXiv.2303.04222. URL: http://arxiv.org/abs/2303.04222.

[42] Krause, Ben et al. *GeDi: Generative Discriminator Guided Sequence Generation*. 2020. arXiv: 2009.06367 [cs.CL].

[43] Kullback, Solomon and Leibler, Richard A. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[44] Lewis, Mike et al. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. arXiv: 1910.13461 [cs.CL].

[45] Li, Zheng et al. *DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization*. 2022. arXiv: 2203.11239 [cs.CL].

[46] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[47] Liu, Chia-Wei et al. "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation". In: *arXiv preprint arXiv:1603.08023* (2016).

[48] Madaan, Aman et al. *Politeness Transfer: A Tag and Generate Approach*. 2020. arXiv: 2004.14257 [cs.CL].

[49] Marie, Benjamin, Fujita, Atsushi, and Rubino, Raphael. "Scientific Credibility of Machine Translation Research: A Meta-Evaluation of 769 Papers". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 7297–7306. DOI: 10.18653/v1/2021.acl-long.566. URL: https://aclanthology.org/2021.acl-long.566.

[50] Mathew, Binny et al. "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 14867–14875.

[51] Mathew, Binny et al. "Thou shalt not hate: Countering online hate speech". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 13. 2019, pp. 369–380.

[52] Mikolov, Tomas et al. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781 [cs.CL].

[53] Min, Bonan et al. *Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey*. 2021. arXiv: 2111.01243 [cs.CL].

[54] Morris, John X. et al. *TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP*. 2020. arXiv: 2005.05909 [cs.CL].

[55] Müller, Karsten and Schwarz, Carlo. "Fanning the Flames of Hate: Social Media and Hate Crime". In: *Journal of the European Economic Association* 19.4 (2020), pp. 2131–2167. ISSN: 1542-4766. DOI: 10.1093/jeea/jvaa045.

[56] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[57] Papineni, Kishore et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[58] Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[59] Peters, Matthew E. et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[60] Qian, Jing et al. "A benchmark dataset for learning to intervene in online hate speech". In: *arXiv preprint arXiv:1909.04251* (2019).

[61] Radford, Alec et al. "Improving language understanding with unsupervised learning". In: (2018).

[62] Radford, Alec et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[63] Ruder, Sebastian. *Recent Advances in Language Model Fine-tuning*. http://ruder.io/recent-advances-lm-fine-tuning. 2021.

[64] Saha, Punyajoy et al. "CounterGeDi: A controllable approach to generate polite, detoxified and emotional counterspeech". In: *arXiv preprint arXiv:2205.04304* (2022).

[65] Sai, Ananya B, Mohankumar, Akash Kumar, and Khapra, Mitesh M. "A survey of evaluation metrics used for NLG systems". In: *ACM Computing Surveys (CSUR)* 55.2 (2022), pp. 1–39.

[66] Saravia, Elvis et al. "Carer: Contextualized affect representations for emotion recognition". In: *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2018, pp. 3687–3697.

[67] Sellam, Thibault, Das, Dipanjan, and Parikh, Ankur P. *BLEURT: Learning Robust Metrics for Text Generation*. 2020. arXiv: 2004.04696 [cs.CL].

[68] Srivastava, Rupesh Kumar, Greff, Klaus, and Schmidhuber, Jürgen. *Training Very Deep Networks*. 2015. arXiv: 1507.06228 [cs.LG].

[69] Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by Ghahramani, Z. et al. Vol. 27. Curran Associates, Inc., 2014. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

[70] Tao, Chongyang et al. "Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.

[71] Tekiroglu, Serra Sinem, Chung, Yi-Ling, and Guerini, Marco. "Generating counter narratives against online hate speech: Data and strategies". In: *arXiv preprint arXiv:2004.04216* (2020).

[72] TensorFlow. *Neural machine translation with a Transformer and Keras*. 2023. URL: https://www.tensorflow.org/text/tutorials/transformer (visited on 06/01/2023).

[73] Thoppilan, Romal et al. "Lamda: Language models for dialog applications". In: *arXiv preprint arXiv:2201.08239* (2022).

[74] Vaswani, Ashish et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL].

[75] Vidgen, Bertie, Margetts, Helen, and Harris, Alex. "How much online abuse is there". In: *Alan Turing Institute* 11 (2019).

[76] Wang, Alex et al. "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: In the Proceedings of ICLR. 2019.

[77] Wang, Xuewei et al. "Persuasion for good: Towards a personalized persuasive dialogue system for social good". In: *arXiv preprint arXiv:1906.06725* (2019).

[78] Warstadt, Alex, Singh, Amanpreet, and Bowman, Samuel R. "Neural Network Acceptability Judgments". In: *arXiv preprint arXiv:1805.12471* (2018).

[79] Warstadt, Alex, Singh, Amanpreet, and Bowman, Samuel R. *Neural Network Acceptability Judgments*. 2019. arXiv: 1805.12471 [cs.CL].

[80] Wilson, Richard Ashby and Land, Molly K. "Hate speech on social media: Content moderation in context". In: *Conn. L. Rev.* 52 (2020), p. 1029.

[81] Wright, Lucas et al. "Vectors for counterspeech on twitter". In: *Proceedings of the first workshop on abusive language online*. 2017, pp. 57–62.

[82] Yang, Zonglin et al. *Logical Reasoning over Natural Language as Knowledge Representation: A Survey*. 2023. arXiv: 2303.12023 [cs.CL].

[83] Yeh, Yi-Ting, Eskenazi, Maxine, and Mehri, Shikib. "A comprehensive assessment of dialog evaluation metrics". In: *arXiv preprint arXiv:2106.03706* (2021).

[84] Zellers, Rowan et al. *Defending Against Neural Fake News*. 2020. arXiv: 1905.12616 [cs.CL].

[85] Zhang, Kelly and Bowman, Samuel. "Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 359–361. DOI: 10.18653/v1/W18-5448. URL: https://aclanthology.org/W18-5448.

[86] Zhang, Tianyi et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL].

[87] Zhang, Weizhe et al. "A transformer–based approach for improving app review response generation". In: *Software: Practice and Experience* (2022). ISSN: 0038-0644. DOI: 10.1002/spe.3158.

[88] Zhou, Chunting et al. "Detecting hallucinated content in conditional neural sequence generation". In: *arXiv preprint arXiv:2011.02593* (2020).

[89] Zhu, Wanzheng and Bhat, Suma. "Generate, Prune, Select: A Pipeline for Counterspeech Generation against 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 (ACL-IJCNLP): Findings*. 2021.

[90] Zhu, Yukun et al. "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 19–27.