Technical University of Munich
Department of Informatics
Research Group Social Computing
Prof. Dr. Georg Groh

Master's Thesis

Countering Sexist Hate Speech

A Transformer-Based Approach for Content Relevant Counter Speech Generation

Author: Yen-Yu Chang
Advisor: Dr. Daryna Dementieva
Supervisor: Prof. Dr. Georg Groh

Submission Date: 17.07.2023

Abstract

Recently research has started focusing on utilizing counter speech in combating online hate on social media platforms, since counter speech avoids undesired effects from content moderation, such as censorship and overblocking. The idea of counter speech is to actively engage in the controversial conversation and intervene directly in the discussion with textual responses that are meant to counter the hate content and prevent it from further spreading. Considering repeated malicious use of neural language models in generating hate speech, manual creation of counter speech is not scalable. Therefore, researchers have put attentions on investigating automatic generation strategies for counter speech. However, this task is relatively new and several existing datasets and generative approaches only tackles general toxic and aggressive hate speech, while generating highly content relevant counter speech for specific target group is rarely investigated. Being aware of the aforementioned challenge, we propose a data creation pipeline for collecting target specific hate and counter speech pairs and plan to apply this pipeline on creating a hate and counter speech pairs dataset for sexist hate speech. Further, we propose a Transformer-based generative language model that utilizes category and hatefulness intensity information as additional features to generate content relevant counter speech for sexist hate speech.

Contents

Chapter 1

Introduction

1.1 Statement of the Problem

Hate speech is an increasing issue in the modern days, as the social interactions are more and more heavily shifted to social media platforms like Twitters, Reddit etc. By leveraging the use of internet platforms, hateful messages can reach a far broader audience with little effort and real-life consequences, as users can hide their identity behind the social media accounts [1]. Hate speech not only causes psychological and emotional harm on the targeted individuals, it can also lead to violent crimes [2] [3].

While censorship has been one of the most common methods to control hate speech [4], many researches suggest that counter speech is a more effective solution to combating online hate [5]. The advantages of counter speech are: (1) it doesn't suppress the public free speech as opposed to censorship [6]; (2) Counter speech encourages other platform users to actively engage in the conversation [7].

Consider the landscape of utilizing counter speech in social media, most counter speech in social conversation are handwritten by human users. While this ensures the quality of produced counter speech, the manual creation of counter speech is not scalable [5]. This issue is further intensified by the malicious use of hate-speech generation language models capable of creating and posting thousands of hateful social media posts in matters of minutes [8].

1.2 Research Question

In the recent years, many researches have introduced methods on generating counter speech to aggressive and toxic languages in general, the topic of gender-specific hate speech has been rarely examined. As Zhu and Bhat [9] stated, the key to effective counter narrative generation is its relevance to the input hate speech. Language models trained on corpora with various topics might be able to generate counter speech to gender-specific hate speech, the outputs tend to be more generic and safe (e.g. "Please refrain from using such language."). This type of generation outputs are acceptable, however, they are rather ineffective in combating hate and supporting the victim demographics.

In this research, we want to address this current lacking of methods that generate counter speech to gender-specific hate speech. We aim to find an approach which can generate **context relevant** counter speech for the given hate speech. Thus, we formulate our research question as:

Q: How to generate context-relevant counter speech against gender-specific hate speech?

We propose to create a data set with hate and counter speech pairs. The hate speech should be categorized according to the form of hate and the intensity of hatefulness should be rated. We propose a transformer-based framework that utilize the category and the emotional intensity of the given hate speech to generate context relevant counter speech, which we will further introduce in the method chapter. This approach is inspired by the work of Zhang et al. [10].

1.3 Ethical Considerations

ss

1.4 Thesis Structure

ss

Chapter 2

Technical Background

2.1 Online Hate

2.2 Countering Hate Speech

2.3 Natural Language Models

2.4 Natural Language Generation

2.5 Evaluation Metrics

Chapter 3

Related Works

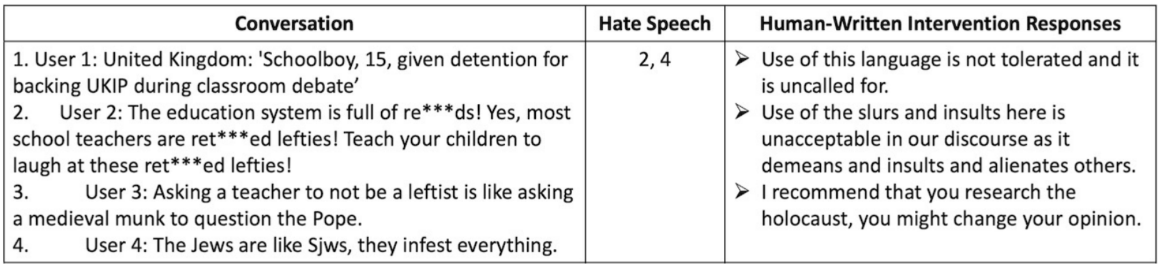
For our research, we focus on three areas that are relevant for counter speech generation: (i) Hate and counter speech datasets, (ii) Counter speech generation methods and (iii) Evaluation metrics for computer generated texts.

3.1 Hate and Counter Speech Datasets

Aiming to encourage strategies of countering online hate speech, a few labeled datasets containing hate and counter speech pairs were introduced.

3.1.1 Benchmark Datasets by Qian et al.

Qian et al. [11] introduced two benchmark datasets containing fully-labeled conversational segments with human-written intervention responses. The two datasets are collected from Gab and Reddit, while the labels and intervention responses are manually created by Mechanical Turk workers via crowd-sourcing. Figure 3.1 shows an example conversation segment.



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 is the conversation text. Indexes are added to each post. Indentations before each post indicate the structure of replies. The second column is the indexes of the human-labeled hateful post. Each bullet point in the third column is a human-written response [11].

These datasets have proven to be useful for training generative models for counter speech generation. Apart from the own work from Qian et al. [11], which experimented with **Seq2Seq**, **Variational Auto-Encoder (VAE)** and **Reinforcement Learning (RL)** to generate hate speech intervention, further studies by other researchers based on these datasets have been conducted [9] [12].

3.1.2 CONAN: Counter-Narratives Datasets to Fight Hate Speech

The CONAN datasets contain four curated datasets for fighting online hate speech. Chung et al. [13] introduced the first part of the four datasets Chung et al. [14], the **CONAN dataset**, which contains multilingual and expert-based hate and counter speech pairs in English, French and Italian, focused on Islamophobia. Figure 3.2 shows an example hate and counter speech pairs in three languages.

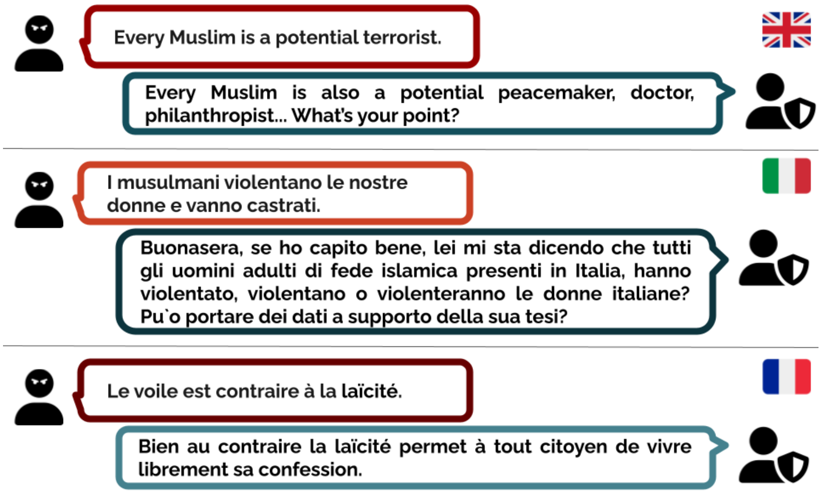
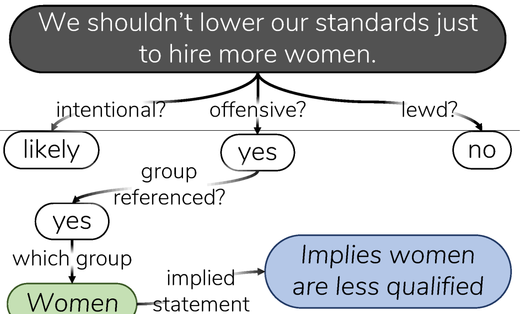

Every Muslim is a potential terrorist.
Every Muslim is also a potential peacemaker, doctor, philanthropist... What's your point?
I musulmani violentano le nostre donne e vanno castrati.
Buonasera, se ho capito bene, lei mi sta dicendo che tutti gli uomini adulti di fede islamica presenti in Italia, hanno violentato, violentano o violenteranno le donne italiane? Pu`o portare dei dati a supporto della sua tesi?
Le voile est contraire à la laïcité.
Bien au contraire la laïcité permet à tout citoyen de vivre librement sa confession.

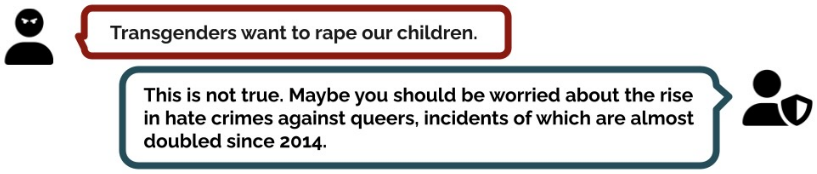
Figure 3.2: An example of the hate and counter speech pairs in the CONAN dataset. One of the three pairs is the originally collected source data and the other two are augmented through translation [13].

The CONAN datasets are expanded in the past years by Fanton et al. [15] (**Multi-target CONAN**), Chung et al. [14] (**Knowledge grounded hate countering dataset**) and Bonaldi et al. [16] (**DIALOCONAN**). Each of the three datasets provided data with different aspects of hate and counter speech, which enable further developments on novel counter narrative generation methods.

The **Multi-target-CONAN** contains 5003 hate and counter speech pairs for English language, covering multiple hate targets (DISABLED, JEWS, LGBT+, MIGRANTS, MUSLIMS, PEOPLE OF COLOR and WOMAN). Figure 3.3 shows an example hate and counter speech pair.

The **DIALOCONAN** focuses on multi-turn hate and counter speech conversation segments (4, 6 or 8 turns). The data are created via a combination of human expert

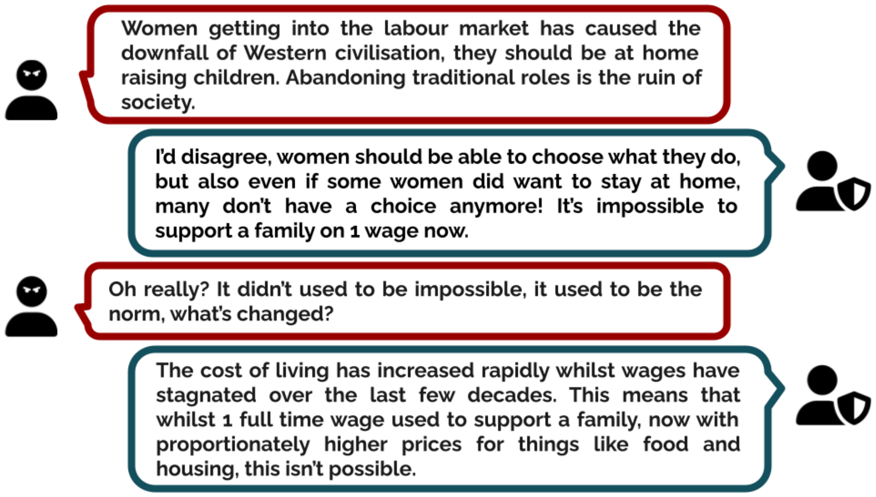




Transgenders want to rape our children.

This is not true. Maybe you should be worried about the rise in hate crimes against queers, incidents of which are almost doubled since 2014.

Figure 3.3: 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 [15].



Women getting into the labour market has caused the downfall of Western civilisation, they should be at home raising children. Abandoning traditional roles is the ruin of society.

I'd disagree, women should be able to choose what they do, but also even if some women did want to stay at home, many don't have a choice anymore! It's impossible to support a family on 1 wage now.

Oh really? It didn't used to be impossible, it used to be the norm, what's changed?

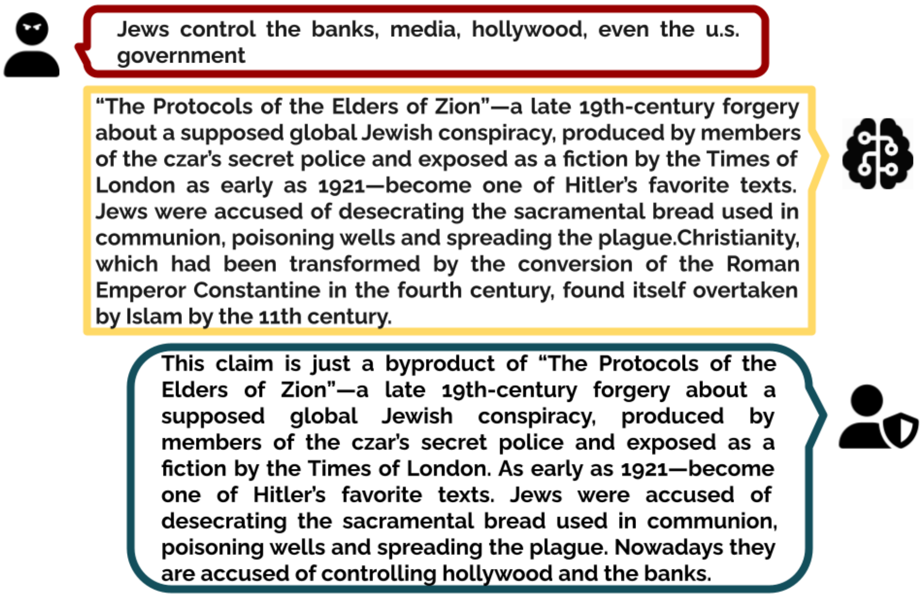
The cost of living has increased rapidly whilst wages have stagnated over the last few decades. This means that whilst 1 full time wage used to support a family, now with proportionately higher prices for things like food and housing, this isn't possible.

Figure 3.4: A 4-turn example of dialogue in DIALOCONAN dataset. The text in red represents posts from a hater and the texts in blue are counter narratives from a NGO operator [16].

intervention over machine generated dialogues, representing conversations between a hater and an NGO operator. The total of 3059 dialogues cover 6 main hate themes (JEWS, LGBT+, MIGRANTS, MUSLIMS, PEOPLE OF COLOR and WOMAN). Figure 3.4 shows an 4-turn dialogue example.

The **Knowledge-Grounded hate countering dataset** focuses on incorporating knowledge into counter speech generation. It contains 195 hate and counter speech pairs, each coupled with the background knowledge used for constructing the counter narrative. Figure 3.6 shows an example.

Besides the datasets containing both the



Jews control the banks, media, hollywood, even the u.s. government

“The Protocols of the Elders of Zion”—a late 19th-century forgery about a supposed global Jewish conspiracy, produced by members of the czar’s secret police and exposed as a fiction by the Times of London as early as 1921—become one of Hitler’s favorite texts. Jews were accused of desecrating the sacramental bread used in communion, poisoning wells and spreading the plague.Christianity, which had been transformed by the conversion of the Roman Emperor Constantine in the fourth century, found itself overtaken by Islam by the 11th century.

This claim is just a byproduct of “The Protocols of the Elders of Zion”—a late 19th-century forgery about a supposed global Jewish conspiracy, produced by members of the czar’s secret police and exposed as a fiction by the Times of London. As early as 1921—become one of Hitler’s favorite texts. Jews were accused of desecrating the sacramental bread used in communion, poisoning wells and spreading the plague. Nowadays they are accused of controlling hollywood and the banks.

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 [14].

hate and counter speech, some other corpora that collect social media posts with hateful messages are also interesting for our research, serving as a starting point for data creation of our own dataset. We briefly introduce two of many, that are especially fitting for our research.

3.1.3 SBIC: Social Bias Inference Corpus

The SBIC is introduced by Sap et al. [17]. It contains 150,000 structured annotations of social media posts, from which 34,000 are biased or hateful towards different demographic groups. Each social media post + annotation pair is called a ”frame”. Figure 3.5 shows two example frames from the corpus.

3.1.4 EDOS: Explainable Detection of Online Sexism

The EDOS dataset[18] is provided for an 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.

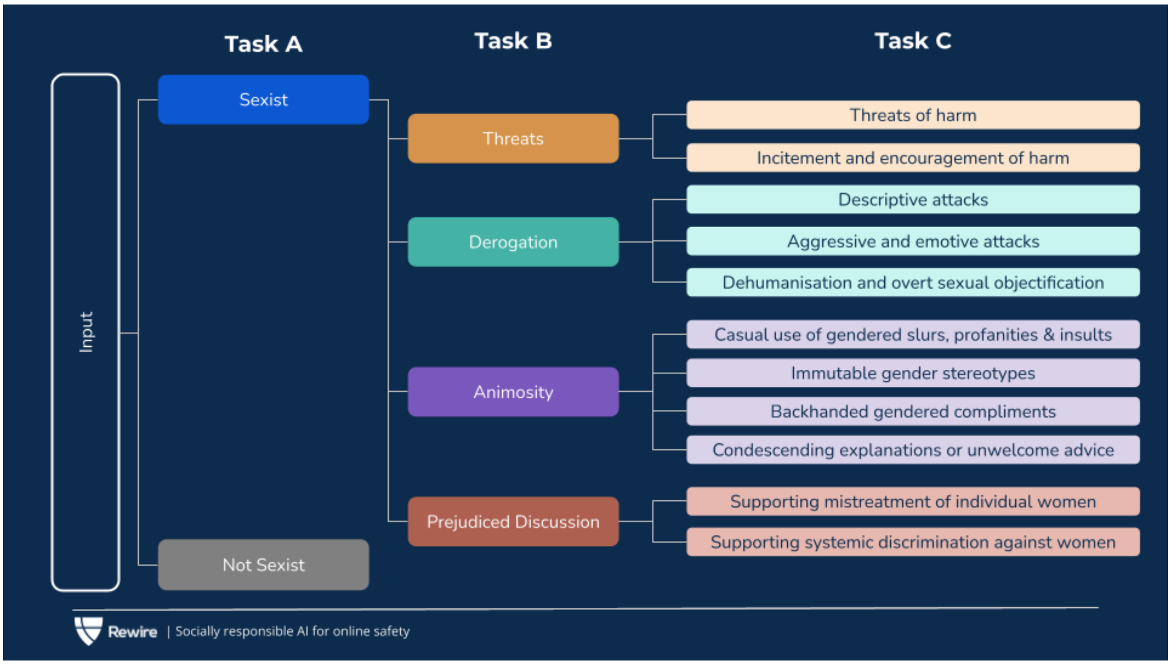

Task A
Task B
Task C
Input
Sexist
Not Sexist
Threats
Derogation
Animosity
Prejudiced Discussion
Threats of harm
Incitement and encouragement of harm
Descriptive attacks
Aggressive and emotive attacks
Dehumanisation and overt sexual objectification
Casual use of gendered slurs, profanities & insults
Immutable gender stereotypes
Backhanded gendered compliments
Condescending explanations or unwelcome advice
Supporting mistreatment of individual women
Supporting systemic discrimination against women
Rewire | Socially responsible AI for online safety

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.

3.2 Counter Speech Generation

There exists a few researches that aim to tackle the task of counter speech generation utilizing language models.

3.2.1 GPT-2

Tekiroglu et al. [19] proposed novel techniques to generate counter speech using a GPT-2 model with postfacto editing by experts or annotator groups.

3.2.2 QIAN

Qian et al. [11] experimented with three methods (Seq2Seq, VAE and RL) to generate counter speech.

3.2.3 VAE

Zhu and Bhat [9] proposed a RNN-based VAE three stages pipeline (Generate, Prune, Select - GPS) to generate diverse and relevant counter speech.

3.2.4 GEDI

Saha et al. [12] proposed an ensemble of generative discriminators (GEDIs) - called CounterGEDI - to guide DialoGPT towards generating more polite, detoxified and emotionally laden counter speech.

3.2.5 Prompt Engineering

Ashida and Komachi [20] investigated the potential of utilizing prompting instead of fine-tuning on large pre-trained language models (GPT-2, textscGPT-Neo and textscGPT-3) to generate counter speech.

3.3 Response Generation

Counter speech generation can be considered as a sub-task of dialogue/response generation. As this task is more intensively researched, we only introduce one publication, from which we drawn our main inspiration.

To automatically generate high quality and relevant responses for app reviews, Zhang et al. [10] proposed a Transformer-based approach, named TRRGen. TRRGen not only uses the review texts as input features, but also utilizes the apps' categories and the review ratings as features. Their results indicate that incorporating these additional features, the model was able to create high-quality responses, outperform state-of-the-art approaches on the task.

Chapter 4

Methods

1. Baseline models:
Vanilla Transformer (TRF)
pre-trained BART
fine-tuned BART without Category embedding
Generate, Prune, Select (GPS)
CounterGEDI

2. Our Approach
pre-trained BART + Category embedding

4.1 Baseline Models

Vanilla Transformer (TRF)
pre-trained BART
fine-tuned BART without Category embedding

4.2 Proposed Model Architecture

Inspired by the work from Zhang et al. [10], we propose a transformer-based generative language model that incorporates the category and the hatefulness of given hate speech as additional features. The architecture is illustrated in figure 4.1.

Category and Intensity Embedding

To incorporate category and intensity information of the hate speech, we propose to define the sexist hate speech into eleven categories according to Kirk et al. [18] and the intensity of hatefulness into ratings from one to five.

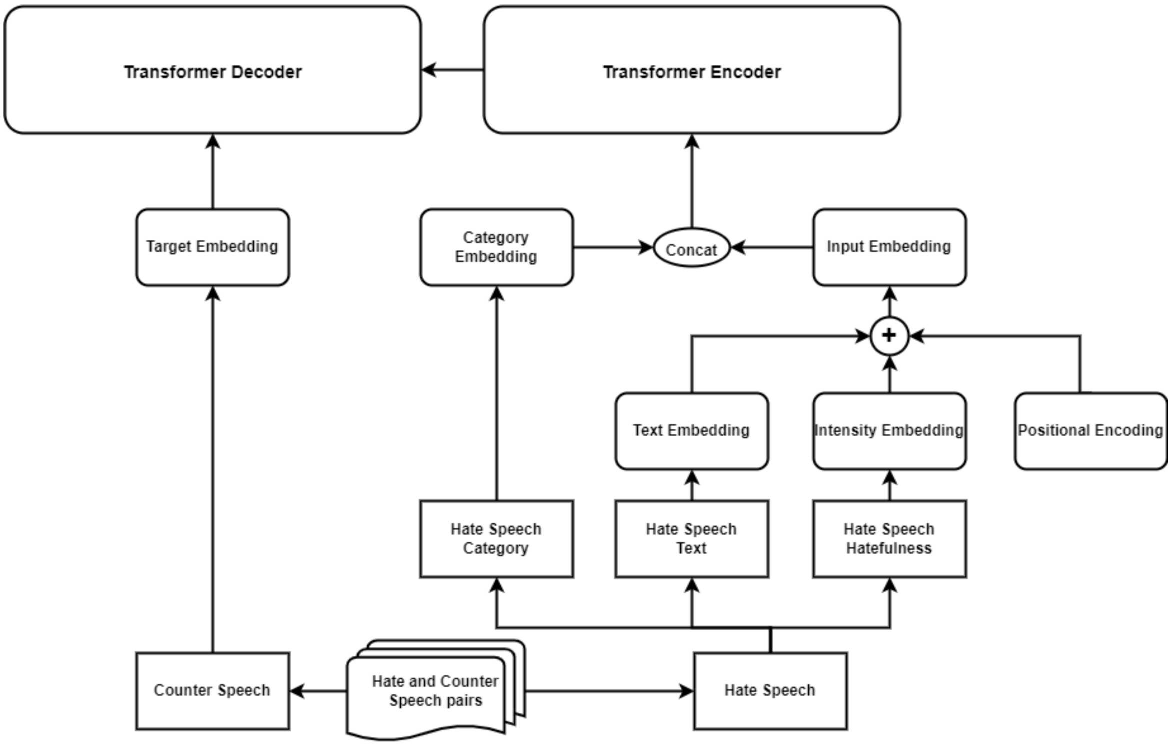


Figure 4.1: Architecture of the proposed transformer-based counter speech generation model.

Each individual category and rating level is added into the vocabulary as distinct new word. Therefore, we can convert them into unique embedding vectors that can be further combined with the word embedding of the hate speech text. Similar to the proposed strategy in Zhang et al. [10], we propose to use addition as the combination operation for the word embedding of hate speech and the intensity embedding and concatenate the result with the category embedding.

4.3 Data Creation and Preparation

To use the existing datasets for our proposed architecture, further data creation pipeline is required, since most datasets contain either only the hate and counter speech pairs without specific hate category and intensity (e.g. the CONAN datasets and the Qian Benchmark dataset), or only the hate speech with their corresponding category (EDOS dataset).

Therefore, we propose to adopt a crowd-sourcing based data creation pipeline introduced by Logacheva et al. [21]. For the corpora containing hate and counter speech pairs, we select those entries related to sexism (target WOMAN) and collect their category. For corpora containing hate speech labeled with specific category, we collect their counter speech.

These two types of datasets can be used as training for crowd-sourcing workers in complement. Workers can train on labeling hate speech category using the hate speech datasets with category labels and train on writing counter speech using datasets containing hate and counter speech pair.

Lastly, all entries require the crowd-sourcing workers to rate their hatefulness intensity.

4.4 Experimental Setup

We propose to compare performances of several existing methods that aim to generate counter speech (e.g. GPS and Counter GEDI), including a baseline model using vanilla Transformer without our proposed category and intensity embedding.

4.5 Data Analysis

We also evaluate models proposed in two related works on our datasets to compare performance:
Generate, Prune, Select (GPS)
CounterGEDI

the statistical or qualitative techniques that were used to analyze the data.

Chapter 5

Results

Chapter 6

Discussion

6.1 Conclusion

discuss the findings in relation to the statement of the problem and the research questions

6.2 Limitations

The limitations section discusses the limitations or weaknesses of the study's design or findings.

6.3 Future Works

Knowledge Grounded Generation - building knowledge base for counter speech + use knowledge base for information retrieval and fact-checking
Style Control - incorporate writing style or tone to further improve the generation quality (sarcastic, calm, empathetic etc.)
Emotional Information - provide further emotional embedding to the model for more content-relevant responses.

Bibliography

[1] C. Bakalis, "Cyberhate: an issue of continued concern for the council of europe's anti-racism commission," 2016. pages 1

[2] Atte Oksanen, Markus Kaakinen, Jaana Minkkinen, Pekka Räsänen, Bernard Enjolras, and Kari Steen-Johnsen, "Perceived societal fear and cyberhate after the november 2015 paris terrorist attacks," *Terrorism and Political Violence*, vol. 32, no. 5, pp. 1047–1066, 2020. pages 1

[3] K. Müller and C. Schwarz, "Fanning the flames of hate: Social media and hate crime," *Journal of the European Economic Association*, vol. 19, no. 4, pp. 2131–2167, 2020. pages 1

[4] A. Álvarez-Benjumea and F. Winter, "Normative change and culture of hate: An experiment in online environments," *European Sociological Review*, vol. 34, no. 3, pp. 223–237, 2018. pages 1

[5] B. Mathew, P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherje, "Thou shalt not hate: Countering online hate speech." [Online]. Available: http://arxiv.org/pdf/1808.04409v2 pages 1

[6] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert, "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech," *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW, 2017. pages 1

[7] J. Garland, K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, and M. Galesic, "Countering hate on social media: Large scale classification of hate and counter speech." [Online]. Available: http://arxiv.org/pdf/2006.01974v3 pages 1

[8] L. Illia, E. Colleoni, and S. Zyglidopoulos, "Ethical implications of text generation in the age of artificial intelligence," *Business Ethics, the Environment & Responsibility*, 2022. pages 1

[9] W. Zhu and S. Bhat, "Generate, prune, select: A pipeline for counterspeech generation against online hate speech." [Online]. Available: http://arxiv.org/pdf/2106.01625v1 pages 1, 5, 8

[10] W. Zhang, W. Gu, C. Gao, and M. R. Lyu, "A transformer–based approach for improving app review response generation," *Software: Practice and Experience*, 2022. pages 2, 9, 10, 11

[11] J. Qian, A. Bethke, Y. Liu, E. Belding, and W. Y. Wang, "A benchmark dataset for learning to intervene in online hate speech." [Online]. Available: http://arxiv.org/pdf/1909.04251v1 pages 4, 8

[12] P. Saha, K. Singh, A. Kumar, B. Mathew, and A. Mukherjee, "Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech." [Online]. Available: http://arxiv.org/pdf/2205.04304v1 pages 5, 9

[13] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, and M. Guerini, "Conan - counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 2819–2829. pages 5

[14] Y.-L. Chung, S. S. Tekiroglu, and M. Guerini, "Towards knowledge-grounded counter narrative generation for hate speech," *arXiv preprint arXiv:2106.11783*, 2021. pages 5, 7

[15] M. Fanton, H. Bonaldi, S. S. Tekiroğlu, and M. Guerini, "Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 3226–3240. pages 5, 6

[16] H. Bonaldi, S. Dellantonio, S. S. Tekiroglu, and M. Guerini, "Human-machine collaboration approaches to build a dialogue dataset for hate speech countering." [Online]. Available: http://arxiv.org/pdf/2211.03433v1 pages 5, 6

[17] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N. A. Smith, and Y. Choi, "Social bias frames: Reasoning about social and power implications of language," *arXiv preprint arXiv:1911.03891*, 2019. pages 5, 7

[18] H. R. Kirk, W. Yin, P. Röttger, and B. Vidgen, "Explainable detection of online sexism (edos)," 2022. [Online]. Available: https://codalab.lisn.upsaclay.fr/competitions/7124#learn_the_details pages 7, 11

[19] S. S. Tekiroglu, Y.-L. Chung, and M. Guerini, "Generating counter narratives against online hate speech: Data and strategies." [Online]. Available: http://arxiv.org/pdf/2004.04216v1 pages 8

[20] M. Ashida and M. Komachi, “Towards automatic generation of messages countering online hate speech and microaggressions,” in *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, and Z. Talat, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 11–23. pages 9

[21] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, and A. Panchenko, “Paradetox: Detoxification with parallel data,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 6804–6818. pages 11