Student's Declaration

I hereby declare that the work presented in the report entitled **Counter Speech Ranking** submitted by me for the partial fulfilment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Engineering* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Md. Shad Akhtar**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

.....*Karan Gupta***.....** **Place & Date:** New Delhi, November 27th, 2024
Karan Gupta 2021258

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date:** New Delhi, November 27th, 2024
Dr. Md. Shad Akhtar

Abstract

This thesis presents the development of a novel ranking method for intent-conditioned counter-speeches. The primary aim of this thesis is to analyze hate speech on online social platforms and identify the most effective strategies to counter it.

The research undertakes an exhaustive analysis of the IntentCONANv2 dataset containing 3488 hate speeches and 13952 counterspeeches across four intents per hate speech. The investigation identifies four parameters mainly- Intensity, Figurative speech, Information presence , and Target group for analysis of the hate speeches. These parameters have further fine-grained definitions to comprehensively analyse hate speech.

This thesis significantly contributes to the field of hate speech and technology by developing a model that effectively analyses hate speech and further assists in the generation of counterspeech.

Keywords: Counter Speech, Hate Speech, Explainable AI, Ranking, Natural Language Processing, Machine Learning, IntenCONANv2

Contents

1 Introduction **2**

1.1 Introduction to Counter Speech 2

1.2 Motivation 3

2 Related Works **4**

3 Dataset **5**

4 Parameters **7**

5 Examples **9**

6 Intent Ranking **12**

7 Basis for Ranking **13**

8 Pipeline **15**

9 Baseline zero-shot **16**

9.1 Methodology 16

9.2 Prompt Design 16

9.3 Zero-Shot Process 17

9.4 Evaluation Metrics and Results 1

Chapter 1

Introduction

1.1 Introduction to Counter Speech

With billions of people interacting on digital platforms daily, the prevalence of hate speech, harassment, and misinformation has escalated, posing significant challenges to individual well-being and societal harmony. Counterspeech refers to the practice of responding to harmful or hateful speech with positive, constructive, or corrective messages. The exposure of social media users to online hate and abuse continues to be a cause for public concern. Prohibiting hate speech on social media, while essential for maintaining respectful and inclusive online spaces, can conflict with freedom of expression, as restrictive measures may suppress legitimate dissent, satire, or controversial opinions, creating a chilling effect where users self-censor Counterspeech has become one of the most effective strategies for combating online hate. It aims to challenge hate, reduce its impact, and promote dialogue without resorting to censorship. [Gupta et al., 2023] in his works introduced a novel way of generating intent-conditioned counterspeech. Fig 1.1 shows some of the different intents defined by [Gupta et al., 2023]. These counterspeeches have better context-specificity and offer a more refined way for moderators to counter hate speeches in online platforms.

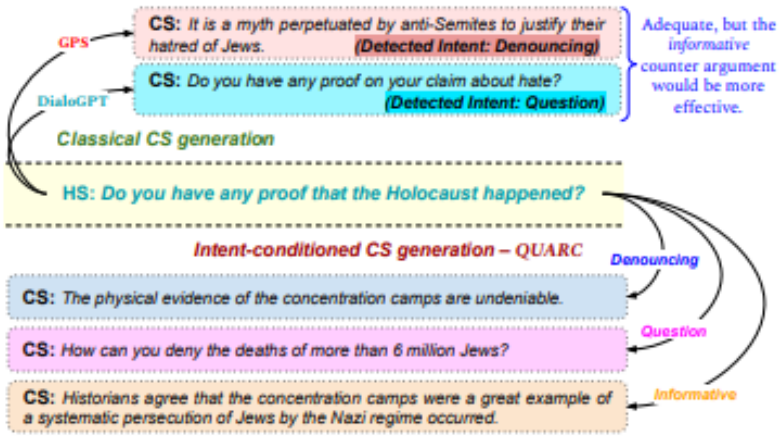


Figure 1.1: Different intents in Hatespeech [Gupta et al., 2023]

1.2 Motivation

[Hengle et al., 2024] in his works introduced a novel framework *CoARL* which analyses the pragmatic implications underlying social biases in hate speech and generates specific intent-conditioned hate speech. The *CoARL* uses Flan-T5 an LLM for the generation of counter-speeches. With the growing cost of generating text from LLMs, generating all the possible counterspeech and then ranking them is impractical in a real-time scenario. We extend upon their work and introduce a ranking method for generating the best intent-conditioned counterspeech given a particular hate speech. Our study introduces several parameters to access hate speech and identify the best intent that would de-escalate the hate speech. We develop mechanism to numerically analyse each parameter and use the analysis to devise a novel ranking mechanism.

Chapter 2

Related Works

1. Intent Conditioned Counterspeech Generation

Automated counterspeech generation has been gaining traction as a viable approach to combating hate speech online. [Gupta et al., 2023] proposed **QUARC**, a two-phase framework for intent-conditioned counterspeech generation, leveraging a novel IntentCONAN dataset comprising 6,831 counterspeech instances across five intents (informative, denouncing, question, humor, and positive). The framework includes a codebook learning module, **CLIME**, to capture intent-specific representations and a conditioned counterspeech generation module, **COGENT**, for generating contextually aligned responses. QUARC significantly outperformed state-of-the-art baselines across metrics such as BERTScore, semantic similarity, and intent accuracy, achieving $\tilde{1}0\%$ improvement in syntactic and semantic evaluation metrics while maintaining diversity in counterspeech outputs.

2. Intent-conditioned and Non-toxic Counterspeech Generation using Multi-Task Instruction Tuning with RLAIF

[Hengle et al., 2024] introduced **CoARL**, a three-phase framework addressing the challenges of short and contextually implicit hate speech. CoARL enhances counterspeech generation by integrating pragmatic explanations of hate speech in its multi-task learning setup, employing task-specific low-rank adapters, and utilizing reinforcement learning with a composite reward function to optimize counterspeech for non-toxicity, effectiveness, and intent alignment. The model's effectiveness was validated using the **IntentCONANv2** dataset, which expanded on prior datasets with 13,952 counterspeech instances across four intents (informative, denouncing, question, and positive). CoARL achieved superior performance in intent conformity and argument quality, demonstrating its ability to generate nuanced and effective counterspeech.

3. Measuring Hate Speech Corpus

The paper introduces the Measuring Hate Speech Corpus dataset, which is highly relevant for analyzing and measuring toxicity in hate speech, especially for our study. The dataset includes labels for critical dimensions like **insult, humiliation, and dehumanization**, which align directly with our parameters for measuring toxicity.[Sachdeva et al., 2022]

4. Fact-Checking Complex Claims with Program-Guided Reasoning

The paper introduces Program-Guided Fact-Checking (ProgramFC), a novel model designed to tackle complex fact-checking tasks by breaking down claims into simpler sub-tasks and using a shared library of specialized functions for resolution. The approach involves two main steps: Reasoning Program Generation and Program Execution This method enhances both explainability and data efficiency and offers clear reasoning paths. [Pan et al., 2023]

Chapter 3

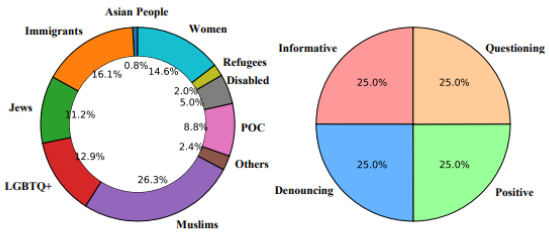
Dataset

We utilize the IntentCONANv2 dataset - a refined version of the publicly available IntentCONAN [Gupta et al., 2023]. The IntentCONANv2 has 3488 Hatespeeches and their 13952 Counterspeech Instances across four intents: Denouncing(DEN), Questioning(QUE), Positive(POS) and Informative(INF). It addresses some key limitations in IntentCONAN - firstly, it removes humour intent due to its subjective nature and tendency to produce vague and offensive content. Secondly, they removed instances where a hate-speech did not have all four types of counterspeech. This ensured an average of four counterspeeches per hate speech. The definitions of the four types are as follows:

- Informative Counterspeech: Uses factual statements to debunk or fact-check claims made by the attacker, making it a widely used form of counterspeech.
- Questioning Counterspeech: Challenging hate speech through questions that compel the speaker to justify or reconsider their statements, supported by factual information when needed.
- Denouncing Counterspeech: Carefully rejects hateful opinions without resorting to name-calling or profanity, ensuring the response remains respectful.
- Positive counterspeech: Uses empathy and civility to reduce online animosity, regardless of the hate speech's severity. Courteous responses can foster a sense of support and humanity, especially for bystanders from the targeted group.

Hate speech	Counterspeech Intent				
Target group	**INF**	**POS**	**QUE**	**DEN**	**Total**
Muslims	914	914	914	914	3656
Women	508	508	508	508	2032
LGBTQ+	449	449	449	449	1796
Jews	392	392	392	392	1568
Refugees	70	70	70	70	280
Asian people	29	29	29	29	116
Immigrants	562	562	562	562	2248
Disabled	173	173	173	173	304
POC	306	306	306	306	80
Other	85	85	85	85	208
Total	3488	3488	3488	3488	13952
Train	2383	2383	2383	2383	9532
Dev	365	365	365	365	1460
Test	740	740	740	740	2960

(a) IntentCONANv2 CS distribution



(a) HS Target distribution (b) CS Intent distribution

(b) IntentCONANv2 distribution

Figure 3.1: Dataset Statistics

Hatespeech	**Informative**	**Questioning**	**Denouncing**	**Positive**
Juice Jews are worse than nukes. Caption to a meme showing Hiroshima 1945 and present day, contrast against pictures of Detroit in 1945 and present day.	Anti-semitism is a serious problem that we need to address and stop spreading relentless hate against people who are innocent.	Want to try a nuke then?	Really? i'd sit with a group of jews any day.	We need to understand from where this hate stems from. stop hating, start loving. the human race needs to be more appreciative of one another.
Israel, Zionism and Jewish Supremacist control of America are the main causes of terrorism, instability and conflict in the world. All other theories are diversions from this ultimate truth.	Wouldn't you say its colonialism from many different countries and the consequences of western intervention abroad which led to a lot of radicalisation and extremism?	Where are the facts?	Alternatively, i think that the main causes of radicalisation and extremism are colonialism and the many western military interventions abroad.	I believe that pointing fingers and hating without cause is not the right way to go about things. this who dun it attitude will not abide if we want to be a peaceful race.

Table 3.1: Examples from the dataset IntentCONANv2

Chapter 4

Parameters

To analyse hate speech, we define the following four parameters:-

1. **Intensity**: **Intensity** measures the severity, strength, and potential harm conveyed by a hateful message. It considers the extent to which the speech humiliates or dehumanizes its target, the level of toxicity in its language, its capacity to incite further hostility or violence, and its exploitation of culturally, historically, or politically sensitive topics. Intensity is measured using three sub-parameters:

a. **Toxicity**: The degree to which the hate speech diminishes the dignity, humanity, or worth of the targeted group. It includes:

- **Insult**: Direct or indirect language intended to belittle or offend.
- **Humiliation**: Content designed to degrade or shame the target group publicly.
- **Dehumanization**: Comparisons or metaphors that strip the target group of human qualities.
- **Blame**: Assigning responsibility for adversity to a particular group or individual.

b. **Incitement**: Assesses the potential of hate speech to escalate hostility or division. Key indicators include:

- **Emotional Appeal**: Use of fear or anger to evoke strong emotional reactions.
- **Polarizing Language**: Divisive phrasing that pits one group against another.
- **Call to Action**: Explicit or implicit prompts for spreading the message or taking action.

c. **Contextual Sensitivity**: Evaluates the exploitation of culturally, historically, or politically sensitive topics, such as:

- **Historical Trauma**
- **Cultural Taboos**
- **Political Tensions**

2. **Figurative Speech**: **Figurative Speech** refers to an implicit or symbolic language, often employing humour, sarcasm, irony, metaphors, or euphemisms. These forms are characterized by higher complexity, masking hateful intent and requiring interpretation to uncover the underlying hostility. This parameter is treated as binary (present/absent) in this research.

3.

Information Presence : **Information Presence** refers to the inclusion of statistical data, numerical claims, or fact-like statements intended to bolster the perceived credibility of the message. Such content increases the message's persuasiveness but also allows for fact-checking.

4. **Target Group** : **Target Group** identifies the individuals, groups, or communities subjected to hostility or derogation, either explicitly or implicitly. The specificity and recognisability of the target group shape the scope and impact of hate speech and guide the appropriate counterspeech responses.

Chapter 5

Examples

1. Intensity :

a. Toxicity

Level	Example	Reasoning
Low	Foreigners on UK benefits leaps 41 in 5 years with over 400K now handed payouts. This nonsense has got to stop.	mildy toxic, little blame, no severe insult
Medium	Israel, Zionism and Jewish Supremacist control of America are the main causes of terrorism, instability, and conflict in the world. All other theories are diversions from this ultimate truth.	Insult, Humiliation and Blame for Terrorism, conflict.
High	When you animals have ruined the world and are eating each other, because that's all you are good for, we'll still be here, keeping our bloodlines pure against all the odds, regretting the day we heeded the womenfolk and took our heels off your throats.	Explicit Dehumanisation

Table 5.1: Toxicity comparison

b. Incitement

Level	**Example**	**Reasoning**
Low	Must be hard being gay. They have to do everything we do but also find the time to molest children.	Toxic but not direct incitement to action
Medium	The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77.	Blaming a community for past traumatic events increasing the risk of escalation
High	Muslims have a bad reputation and there will be a reason. I even saw a video where they kill dogs with acid!	Emotional Appeal to incite hostility.

Table 5.2: Incitement Comparison

c. contextual sensitivity

Level	Example	Reasoning
Low	Minorities vote to strip wealth and power from White people and they are organised to do this by the Jews.	General Claims but not culturally sensitive.
Medium	Jews cannot be patriots, since their allegiance will always be to the state of Israel.	Understandable without context but still carries meaning.
High	What's the difference between a Cow and the Holocaust? You stop milking the cow after 70 years.	undermines the holocaust as a traumatic experience for the Jews.

Table 5.3: Contextual Sensitivity

2. Info Presence

Level	Example	Reasoning
Low	I tolerated fag pride. I tolerated fags on TV and in every town square. I draw the line at tranny demons reading books to children.	No stats or verifiable claims
Medium	Foreigners on UK benefits leaps 41 in 5 years with over 400K now handed payouts. This nonsense has got to stop.	Numerical Claims
High	Pakistan 202 Million Nigeria 193 Million Bangladesh 165 Million... Great Britain 66 Million.	Detailed numerical claims

Table 5.4: Info Presence

3. Target Group

Level	Example	Reasoning
Not Present	For all the people who believe that without migration, the NHS cannot do it, I would like to tell you that we have already done it in history! precisely before the 90s and before Tony Blair	no specific target - just migrants
Implicit	When you animals have ruined the world and are eating each other, because that's all you are good for, well still be here, keeping our bloodlines pure against all the odds, regretting the day we heeded the womenfolk and took our heels off your throats.	Targeted at women but implicitly.
Explicit	The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77.	Specially targets the Jew community.

Table 5.5: Target Group

4. Figurative Speech

Level	Example	Reasoning
YES	The British people should use those houses, not immigrants.	No use of sarcasm, humour, metaphors - easily understandable
NO	Race is a social construct, and if you see any difference between these two women, it's because you have a mental illness. image showing a blonde, European female athlete, juxtaposed against an image of an African woman wearing a lip plate	Use of sarcasm - what is being implied is opposite of what is written.

Table 5.6: Figurative Speech

Chapter 6

Intent Ranking

		P	Q	I	D
Intensity	High	3	4	2	1
	Medium	1	2	3	4
	Low	2	1	3	4
Info Presence	High and misinformation	3	2	1	4
	High and correct info	1	2	4	3
	Low	3	2	1	4
Figurative	NO				
	YES				
Target Group	Explicit	1	3	4	2
	Implicit	2	1	3	4
	Not Present	1	2	3	4

Table 6.1: Ranking of Counterspeech Intents Based on Parameters.
P: Positive, Q: Questioning, I: Informative, D: Denouncing

Chapter 7

Basis for Ranking

Counterspeakers aim to target bystanders on social media platforms. These bystanders are not (at least yet) generating hateful language themselves, but rather are people exposed to hateful content either incidentally or by active engagement. Here, counter speakers hope to persuade bystanders that the hateful content is wrong or unacceptable, again by deconstructing and delegitimising the hateful narrative. The strategy here may be to offer facts, point out hypocrisy, denounce the content, or use humour to discredit the speaker.[Chung et al., 2023]

Information Presence
Factual counter-speech delivers facts and supports the target group. Factual counter speech is shown to contribute to a deliberative discussion atmosphere (e.g. mutual respect, openness for different views) increasing the willingness of others to partake.[Buerger, 2021] In case of misinformation, providing the correct information helps in countering false or partially true hateful comments. Counterspeakers often aim their messages at the broader audience, particularly the "movable middle," rather than directly confronting those spreading hateful speech or misinformation. This strategy recognizes that hateful content spreads quickly and widely. Many counter speakers choose to engage when they have expertise on the topic, using factual information to document dissent and provide credible counterpoints.[Buerger, 2021]
While research suggests that fact-checking rarely changes entrenched opinions, it can influence silent readers—those who follow the discussion without participating. These readers may gain the confidence to reject misinformation or even become counterspeakers themselves. By sharing credible articles and presenting counterarguments, counter speakers hope to persuade moderate individuals and inoculate them against hateful narratives, even if they are unlikely to sway extremists. [Buerger, 2021]

Denouncing
In this strategy, counterspeakers denounce the message as being hateful. This strategy can help the counterspeakers reduce the impact of the hate message.[Mathew et al., 2019] A second key way denouncing functions is to show support directly to targets of hate. Online abuse can harm targets' psychological well-being, causing fear, threats, and concerns for safety. Counterspeakers help mitigate these effects by challenging abuse, offering support to targets, and encouraging bystanders to do the same. This support reassures targets that they are not alone and that the perpetrator's attitudes are not widely shared. Strategies include denouncing hate and expressing positive sentiments toward the targeted group, fostering intergroup solidarity, and potentially reducing retaliatory antagonism.[Chung et al., 2023]

Positive
Empathy-based counterspeech can consistently reduce hate speech, although this effect is small. The study by [Hangartner et al., 2021] investigates the impact of three counterspeech strategies on reducing xenophobic and racist hate speech on Twitter through a field experiment. The researchers identified xenophobic tweets using a combination of dictionary-based approaches, sentiment analysis, and manual annotation. A total of 1,350 users who posted such tweets were randomly assigned to one of three interventions (20% probability each) or a control group (40% probability, no intervention). The interventions involved one-time counterspeech messages sent by neutral human-controlled bots within 24 hours of the original xenophobic tweet.

To measure the effects, the study assessed:

- Removal of past xenophobic tweets.
- Creation of xenophobic tweets and total tweets over a 4-week follow-up.
- Negative sentiment of tweets during follow-up (using the Vader compound score).

The results showed that the empathy-based counterspeech strategy had small but notable effects:

- Reduced xenophobic tweet creation by 0.10 SD and total tweets by 0.13 SD (1.3 fewer xenophobic tweets and 91.6 fewer total tweets on average).
- It increased the likelihood of deleting the original xenophobic tweet by 0.21 SD (8.4 percentage points higher).

These findings suggest that empathy-based counterspeech can modestly reduce hate speech and promote tweet deletions.[Hangartner et al., 2021]

Questioning
Counter-questions can be an effective strategy to address hate speech by prompting reflection and encouraging dialogue. Unlike direct confrontation, counter-questions subtly challenge the speaker's assumptions and biases, fostering critical thinking without escalating conflict. This approach often de-escalates tension, creates opportunities for constructive dialogue, and may influence the speaker to reconsider their stance.

Counter-questions are generally more effective and preferred when the intensity of hate speech is low. In such cases, the speaker is likelier to engage in dialogue rather than react defensively. Low-intensity hate speech often stems from ignorance or thoughtlessness rather than deeply entrenched hostility, making it more amenable to subtle interventions like counter-questions. These questions can encourage the speaker to reflect on their words and consider alternative perspectives, fostering constructive conversation without escalating the situation.

Chapter 8

Pipeline

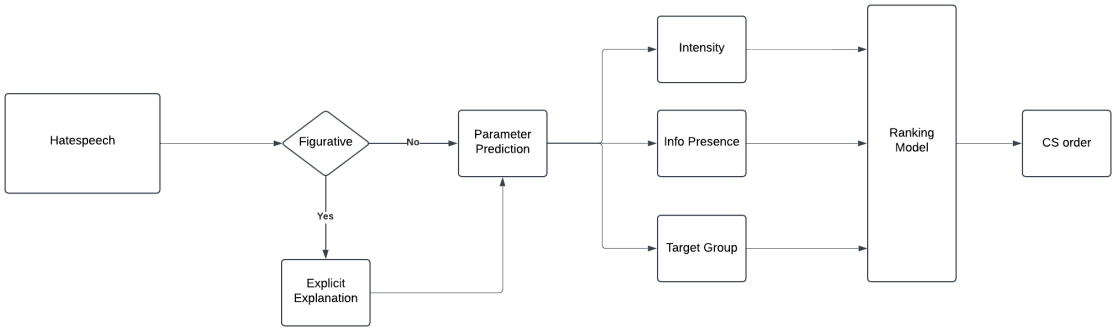


Figure 8.1: Pipeline for Ranking

Figure 7.1 shows the complete pipeline from the hate speech to the ranking of different intents. The steps involved are as follows:

- Step 1 involves checking if the hate speech has some figurative speech and extracting its explicit meaning.
- The speech with the explanation, if any, is passed to the parameter prediction model.
- The model predicts the values of the parameters: Intensity, Information presence and Target group, which are passed to the ranking model.
- The ranking model uses the hate speech and the parameters to rank the intents.

Chapter 9

Baseline zero-shot

9.1 Methodology

We used ChatGPT4o as a zero-shot baseline to annotate hate speech across all four parameters - **Intensity, Figurative Speech, Information Presence, and Target Group.**

9.2 Prompt Design

The following prompt was used to guide GPT4o's evaluation :

Prompt Text:

> **Parameters to Evaluate**:
>
> **Intensity of Hate Speech** *(Rate as Low (1), Medium (2), or High (3))*
>
> - **Definition**: The intensity of hate speech refers to the severity, strength, and potential harm conveyed by a hateful message. It encompasses the extent to which the speech humiliates or dehumanizes its target, i.e., the level of toxicity in its language, its capacity to incite further hostility, violence, or division, and the degree to which it exploits culturally, historically, or politically sensitive topics.
>
> **Figurative Speech** *(Answer Yes or No)*
>
> - **Definition**: Use of implicit, indirect, or symbolic language to convey hateful messages.
> - **Includes**: Humor, Sarcasm, Irony, Metaphors, Euphemisms.
> - **Note**: Characterized by higher complexity where the hateful intent is masked or layered.
>
> **Information Presence** *(Rate as Low (1), Medium (2), or High (3))*
>
> - **Definition**: Inclusion of statistical data, numerical claims, or fact-like statements aiming to bolster the perceived credibility of the message.
> - **Consider**:
>   - The opportunity for verification through fact-checking.

– The presence of references to identifiable events, data points, or specific entities, regardless of accuracy.

Target Group Specificity *(Rate as Not Mentioned (1), Implicitly Mentioned (2), or Explicitly Mentioned (3))*

- **Definition**: Identifiability of the individual, group, or community subjected to hostility.
- **Not Mentioned (1)**: No clear target is identifiable.
- **Implicitly Mentioned (2)**: Target is suggested indirectly or requires inference.
- **Explicitly Mentioned (3)**: Target is clearly and directly identified.

Instructions:

1. **Read** the provided text carefully.
2. **Evaluate** each parameter separately based on the definitions and criteria.
3. **Provide** the rating for each parameter along with a brief justification citing specific elements from the text that influenced your assessment.
4. **Format** your response as follows.

Format:

1. Intensity of Hate Speech: [Rating (1-3)]
   - Justification: [Provide reasons based on Toxicity, Incitement, and Contextual Sensitivity.]
2. Figurative Speech: [Yes/No]
   - Justification: [Explain whether figurative language is used and how.]
3. Information Presence: [Rating (1-3)]
   - Justification: [Indicate the presence of data, statistics, or factual claims.]
4. Target Group Specificity: [Rating (1-3)]
   - Justification: [Describe how the target group is referenced in the text.]

9.3 Zero-Shot Process

- Each piece of hate speech text was provided as input along with the above prompt.
- GPT-4o returned structured outputs containing ratings and justifications for each parameter.

Hatespeech	Parameter	Human Label	ChatGPT label
The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77	Intensity	2	3
The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77	Figurative	NO	NO
The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77	Info Presence	3	3
The Jews did USS Liberty Attack, the King David hotel attack, New York 911, London 77	Target Group	3	3

Table 9.1: Example from our annotation comparison csv

9.4 Evaluation Metrics and Results

We compared GPT4o annotations to the human-curated annotations and evaluated the following metrics

- **Accuracy**: Proportion of correctly predicted labels.
- **Precision**: Correct predictions among all predictions made for a class.
- **Recall**: Correct predictions relative to total true instances of a class.
- **F1-Score**: Harmonic mean of precision and recall.

Parameter	Accuracy	Precision	Recall	F1 Score
Intensity	0.69	0.60	0.69	0.61
Figurative	0.75	0.81	0.75	0.75
Info presence	0.81	0.86	0.81	0.81
Target Group	0.75	0.83	0.75	0.72

Table 9.2: Results

Bibliography

Catherine Buerger. #iamhere: Collective counterspeech and the quest to improve online discourse. *Social Media + Society*, 7(4):20563051211063843, 2021. doi: 10.1177/20563051211063843. URL https://doi.org/10.1177/20563051211063843.

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. Understanding counterspeech for online harm mitigation, 2023. URL https://arxiv.org/abs/2307.04761.

Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation, 2023. URL https://arxiv.org/abs/2305.13776.

Dominik Hangartner, Gloria Gennaro, Sary Alasiri, Nicholas Bahrich, Alexandra Bornhoft, Joseph Boucher, Buket Demirci, Laurenz Derksen, Aldo Hall, Matthias Jochum, Maria Munoz, Marc Richter, Franziska Vogel, Salomé Wittwer, Felix Wüthrich, Fabrizio Gilardi, and Karsten Donnay. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118:e2116310118, 12 2021. doi: 10.1073/pnas.2116310118.

Amey Hengle, Aswini Kumar, Sahajpreet Singh, Anil Bandhakavi, Md Shad Akhtar, and Tanmoy Chakroborty. Intent-conditioned and non-toxic counterspeech generation using multi-task instruction tuning with rlaif, 2024. URL https://arxiv.org/abs/2403.10088.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. Thou shalt not hate: Countering online hate speech, 2019. URL https://arxiv.org/abs/1808.04409.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. Fact-checking complex claims with program-guided reasoning, 2023. URL https://arxiv.org/abs/2305.12744.

Pratik S Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris J Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @ LREC 2022*, pages 83–94. European Language Resources Association (ELRA), 2022.