Distilling Implied Bias from Hate Speech for Counter Narrative Selection

Nami Akazawa[1,2], **Serra Sinem Tekiroğlu**[2], **Marco Guerini**[2],
[1]University of Trento, Italy
[2]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
nakazawa@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu

Abstract

Hate speech is a critical problem in our society and social media platforms are often an amplifier for this phenomenon. Recently the use of Counter Narratives (informative and non-aggressive responses) has been proposed as a viable solution to counter hateful content that goes beyond simple detection-removal strategies. In this paper we present a novel approach along this line of research, which utilizes the implied statement (bias) expressed in the hate speech to retrieve an appropriate counter narrative. To this end, we first trained and tested several LMs that, given a hateful post, generate the underlying bias and the target group. Then, for the counter narrative selection task, we experimented with several methodologies that either use or not use the implied bias during the process. Experiments show that using the target group information allows the system to better focus on relevant content and that implied statement for selecting counter narratives is better than the corresponding standard approach that does not use it. To our knowledge, this is the first attempt to build an automatic selection tool that uses hate speech implied bias to drive Counter Narrative selection.

1 Introduction

When visiting Social Media Platforms (SMPs), it is common to stumble upon content that is hateful or discriminatory. These posts usually address derogatory cliché about a specific community. For instance, coming across the microaggression below (Breitfeller et al., 2019), the reader naturally infers that this post is targeting a specific group of people—Muslims. Most people would additionally recognize the indirectly-stated stereotype, a so-called false narrative, i.e., "*Muslims are terrorists.*". Without prior stereotype familiarity, one would not fully understand the implied meaning emerging from the reading.

"*Wow, don't get the Muslim mad guys! We don't want to come to a blown up school tomorrow!*"

SMPs have become means to spread this kind of inappropriate content about target groups (TG). Corporations owning social media platforms counter this phenomenon through various mechanisms, such as by allowing users to report hateful posts and suspicious accounts. However, manual intervention is a time consuming activity for all parties involved. Developing automated hate speech (HS) detection systems has thus become a research priority. Since the early stage of HS classification done by Warner and Hirschberg (2012), many competitive models and datasets emerged (Malik et al., 2022b; Mathew et al., 2021; Vidgen et al., 2020; Fortuna and Nunes, 2018). However, these solutions are not sufficient and can be charged with censorship as argued in (Benesch, 2014; Williams, 2019). An alternative and more promising way to tackle online hatred is by responding to HS with non-aggressive textual feedback using credible evidence and factual arguments that give alternative viewpoints (Benesch, 2014; Gagliardone et al., 2015). This strategy, called Counter Narrative (CN), is believed to be effective at moderating the hate content while preserving freedom of speech (Schieb and Preuss, 2016; Silverman et al., 2016; Wright et al., 2017; Ziegele et al., 2018; Ziems et al., 2021). Still, since manually crafting CNs is a time consuming activity, recent research started to draw attention to automatic CN generation task (Qian et al., 2019; Chung et al., 2021a; Tekiroglu et al., 2022; Zhu and Bhat, 2021).

To build an automatic CN generation tool, it is vital for the system to understand the underlying meaning of the HS, and not to rely solely on strongly indicative keywords, such as explicit slurs (Yin and Zubiaga, 2022). In this regard, pragmatic inference is required to understand the complete meaning of HS, usually built upon stereotype

Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA), pages 29–43

TG	HS	IS	CN
MIGRANTS	Immigration is a curse for Britain and the world. It is destroying our civilization.	Immigration is bad.	Some countries are importing a lot more people than others. Our country is wealthy, I are a thriving place to live. I should not blame immigrants.
LGBT+	Gays are worse that nukes. They like to force their beliefs on others. They are gaslighting us!	LGBT+ have an agenda.	Well, since you are very scared, I am sure you have some examples on when some gay tried to convince you to join their side. I would be interested in hearing that.
WOMEN	Women must not have equal rights, they are merely a burden to society, since they want to steal men's rights.	Women are inferior.	Equality of rights is the most fundamental human concept, and feminism is about eliminating inequality, not removing rights from men.

Table 1: Examples of diverse HSs with hate target, the corresponding IS and a suitable counter-narrative.

knowledge, e.g., "*Asians are good at math*". Since multiple HS could map to the same implied statement (IS), our hypothesis is that narrowing down HS to clear and accurate IS makes it easier to find a proper CN for a given HS. Table 1 shows examples of different HS discourses with corresponding IS and a possible CN.

In this work we present two different, sequential tasks. The first consists of distilling/generating implied statement from hate speech utterances and it is achieved by fine-tuning and comparing different pre-trained neural language models (LMs). The second consists of testing various methodologies based on semantic similarity that, given a hateful post and an implied statement, find the most suitable CN among the entries of a dataset comprising HS–CN pairs (Fanton et al., 2021)

2 Related Work

In this section we briefly present NLP approaches to deal with HS, IS, and CN.

Hate Speech detection is used to maintain healthy online environments (Yin and Zubiaga, 2021; Jahan and Oussalah, 2021; Schmidt and Wiegand, 2017; Alkomah and Ma, 2022; Malik et al., 2022a). HS detection was initially cast as a binary classification task (Badjatiya et al., 2017; Fortuna and Nunes, 2018; Nobata et al., 2016); while recently interest shifted on fine-grained categories of HS, such as its type, target, implicitness, rationale (Zampieri et al., 2019; Qian et al., 2018; ElSherief et al., 2021; Sap et al., 2019; Fortuna et al., 2019).

Training examples are usually scraped from online platforms like Reddit, Twitter and Gab (Sap et al., 2019; Mollas et al., 2022; Gao et al., 2017; ElSherief et al., 2021; Waseem and Hovy, 2016; Zampieri et al., 2019; Kennedy et al., 2018). A particular difficulty for the detection task is the generalizability across datasets (Swamy et al., 2019; Caselli et al., 2021).

Implied Statement Generation Recently, a new line of research emerged, focusing on generating the IS that underlies an HS instead of simply assigning a label to it (Sap et al., 2019; ElSherief et al., 2021). In particular, Sap et al. (2019) designed a model generating TGs, ISs and HS characteristics, such as posts' offensiveness. The auto-regressive models they used generated TGs accurately; however, they struggled with generating ISs, especially when the HSs and their implications had low lexical overlap.

ElSherief et al. (2021) focused on implicit HS and defined a six-class taxonomy, including IS. For the generation of labels for each implicit HS, results revealed that GPT-2 with the beam search option performed best compared to the other models. However, the GPT-2 model performed worse for the IS generation task.

Both publications shared common conclusions on the IS generation—it performed worse than the label generation using GPT models due to texts being longer, more subtle, and having low lexical overlap with the HS posts. Nonetheless, this could be partially due to certain degree of subjectivity in the perceived hatred, so that different annotators can come up with different IS wording during data collection (Sap et al., 2022; Rottger et al., 2022). We will also address this hypothesis in our experiments.

Counter Narrative Generation. Writing CNs is time-consuming activity so automatic ways to generate CN are beginning to be investigated. A comparative study by Tekiroglu et al. (2022) showed that among different LMs, the autoregressive models like GPT-2 with stochastic decoding methods achieve the best results. Zhu and Bhat (2021) focused on improving diversity and relevance of the generated CN using a three-stage pipeline com-

posed of a model to generate CN candidates, a BERT classifier to filter ungrammatical CNs, and an information retrieval method to select the most relevant CN. An alternative approach to increase diversity and relevance is to add background knowledge to the CN generation by incorporating factual, up-to-date content retrieved from external resources (Chung et al., 2021a).

Focusing on available datasets, COunter NArratives through Nichesourcing (CONAN) was the first large-scale multilingual dataset with three different languages to fight Islamophobia (Chung et al., 2019). To mitigate the difficulty of nichesourcing, an Author-Reviewer Architecture was proposed by Tekiroglu et al. (2020). Following Tekiroglu et al. (2020) method, Fanton et al. (2021) produced Multi-Target CONAN dataset using a iterative human-in-the-loop approach.

3 Datasets

Three datasets were used for our experiments.

Social Bias Inference Corpus (SBIC) contains roughly 150,000 structured annotations of posts collected from various social media sources (Sap et al., 2019). The corpus is annotated following a formalism (Social Bias Frames) that aims to cover both structured pragmatic and social implications by including various categorical and textual annotations. The authors considered labels informing whether the post was offensive, intentionally offensive, lewd, implicating any group, or using in-group language. Whenever TGs and ISs were available, they were annotated in a free-text format, resulting in about 34,000 detailed implication examples. ISs were generally annotated in the form of simple Hearst-like patterns (e.g., "*immigrants are ⟨ADJ⟩*"; (Hearst, 1992)). Some posts were annotated by multiple humans, thus receiving multiple ISs. SBIC covers a variety of HSs that target different groups, depending on gender, race, and religion.

Implicit Hate Corpus (IHC) ElSherief et al. (2021) focuses on implicit HS, and covers diverse hate language, such as indirect sarcasm, intimidation, and euphemisms. The authors define a theoretically grounded framework with a fine-grained six-class taxonomy to cover different characteristics of implicit HS. In addition to the categories, the dataset includes descriptions of IS and TG labels. The posts are collected mainly from Twitter, focusing on US hate groups as identified by the Southern Poverty Law Center report[1].

Each post got assigned one out of six hate categories, a free-text TGs and ISs formatted as Hearst-like patterns.

Multi-Target CONAN (MT-CONAN) dataset contains HS–CN pairs (Fanton et al., 2021) that are collected through an human-in-the-loop data collection methodology. This dataset was obtained starting from a seed dataset of pairs written by NGO experts and then expanded iteratively using GPT-2 to generate new examples. It covers 5,003 HS/CN pairs in seven TGs: MUSLIMS, JEWS, LGBT+, WOMEN, DISABLED, MIGRANTS, and PEOPLE OF COLOR (POC).

4 Data Preparation

For IS generation and CN selection tasks, data filtering and pre-processing techniques were applied to the datasets we described.

For Implied Statements, the goal of data preparation was to keep combinations of HS, TG, and IS that follow specific patterns and meet specific requirements. For this reason we: (i) selected only hateful sentences that have an annotated IS, (ii) select samples with a single-sentenced IS that rigorously follow the pattern $\langle subject \rangle$ – $\langle predicate \rangle$ – $\langle object \rangle$. (iii) select examples with targets that can be aligned with those in MT-CONAN dataset. The original datasets contain 153,498 samples from IHC and SBIC while after filtering and standardization we obtained a total of 30,585 samples. We call this resulting dataset IMPLIED BIAS STATEMENT DATASET (IBSD). A detailed description of the whole procedure can be found in Appendix A.

For Counter Narratives, instead, the goal of data preparation was simply to exclude from MT-CONAN the examples that were using an HS from SBIC, to grant mutually exclusive data for the CN selection task. This filtering procedure resulted in a total of 4,251 examples.

5 Implied Statement Generation

To generate IS from HS we compared different transformer-based LMs, where each model is trained with HS–IS pairs from IBSD. Additionally, we tested several input options and decoding strategies to find the best setting for generating the IS.

[1] splcenter.org/hate-map

5.1 Input Settings

For the IS generation task we differentiate between HS with multi-ISs and HS with single ISs. We performed a stratified sampling according to these two categories (an HS only appears in training or test set, regardless of being single or multi-IS) and according to TGs using the ratio of 8:1:1 for train, development, and test sets respectively. We then experimented with three different settings:

$$[\texttt{HS}][H_0 : H_M] \rightarrow [\texttt{IS}][I_0 : I_K] \quad (1)$$

$$[\texttt{HS}][H_0 : H_M] \rightarrow [\texttt{TG}][T_0 : T_N][\texttt{IS}][I_0 : I_K] \quad (2)$$

$$[\texttt{TG}][T_0 : T_N][\texttt{HS}][H_0 : H_M] \rightarrow [\texttt{IS}][I_0 : I_K] \quad (3)$$

The special tokens $[HS]$, $[IS]$, $[TG]$ mark the beginning of "Hate Speech Post", "Implied Statement", and "Target Group", followed by HS, IS, and TG sequences, which are indicated as $[H_0 : H_M]$, $[I_0 : I_K]$, and $[T_0 : T_N]$, respectively.

The first two configurations are designed as using only hate speech as input, which represent the most realistic run-time scenario while tackling hate speech. The third configuration, instead, includes the additional TG information in the input (available at training time in our datasets) to investigate how the model could leverage this additional information to generate IS.

5.2 Setup

Following the studies by ElSherief et al. (2021) and Sap et al. (2019), we use GPT-2 for our experiments. Additionally, we tested BART and T5 models considering their effectiveness in summarization and question-answering tasks (Lewis et al., 2019; Raffel et al., 2020). We have experimented with Greedy search, Beam search, Temperature sampling, Top-*k* and Top-*p* as decoding methods. Details of training configurations, hyperparameter settings, and decoding values can be found in Appendix B.

5.3 Results

The implied statement generation results under the three input-output configurations are shown in Table 2, while the results of target generation under the configuration 2 are reported in Table 7 in the appendix section. Since each TG category is a single word, we evaluated the results as a classification task rather than a generation one, consequently we used F1 scores. For the generation experiments, we utilized unigram, bigram-based BLEU, and ROUGE (also ROUGE-L) metrics, as the IS sentences are generally short (average length is 5.41 tokens). In addition, we report repetition rate (RR) scores, which measures word repetitiveness in generated text and is computed by the rate of non-singleton n-grams it contains (Cettolo et al., 2014; Bertoldi et al., 2013).

From the results, three major conclusion can be drawn. (i) *BART outperformed the other models*, yielding higher BLEU and ROUGE scores for all the options. (ii) *Deterministic decoding proved to be the most suitable for this task* under all decoding strategies/model configurations (greedy search and beam search resulted in the highest performance for all models in all options)[2]. (iii) BLEU and ROUGE scores are generally much higher, usually more than twice, for multi-ISs than for the single-referenced IS (e.g., Single 23.1 and Multiple 51.1 for BLEU-2 BART$_{gdy}$ in Table 2). While the trivial explanation is that there is a higher probability that the generated IS has a prominent overlap with a gold IS in the multi-reference setting, this seems also to indicate that the possible degree of subjectivity in perceived ISs (i.e different annotators can come up with different IS wording for the same post) is well captured with few references rather than one. This also indicates that generations can be of higher quality than what the numbers tell: from a manual analysis of small overlapping examples we found that this was often the case, i.e. different IS wording rather than a poor generation quality.

In Configuration 2, all models/decoding methods (except for T5$_{smp/top-k/top-p}$) did well at generating the TGs ($MicroF1 > 87$). BART$_{gdy}$ configuration performed the best for most of the targets and overall F1 in TG classification[3].

For Configuration 3, in which the model generates IS, given TG and HS, we observed similar result patterns to Configuration 1 and 2. BART with greedy search and beam search outperformed all other model/decoding settings. It also scored higher both on BLEU and ROUGE than Configura-

[2]We investigated several different decoding values for stochastic decoding, where top-$p \in \{0.0, 0.1, ..., 1.0\}$, top-$k \in \{10, 20, ..., 100\}$ and temperature $\in \{0.1, 0.2, ..., 1.0\}$. The closer the decoding value corresponds to the greedy search, the higher the BLEU and ROUGE scores. This supports the overall finding in different input options, that the greedy search is preferred.

[3]Considering F1 scores for each target, MIGRANTS and DISABLED scored the lowest (below 90 F1 score). This is likely because both TGs make up a smaller proportion of the IBSD (e.g., DISABLED is 3.1% of the data.)

Models	BLEU-2		ROUGE-1		ROUGE-2		ROUGE-L		RR
	Single	Multiple	Single	Multiple	Single	Multiple	Single	Multiple	
Input/Output Configuration 1									
T5$_{gdy}$	20	48.2	39.6	58.7	20.2	38	39.2	58.5	40
T5$_{beam}$	21.2	48.7	40.1	59.4	20.5	39	39.5	59.2	34.7
T5$_{smp}$	15.8	31.4	32.6	47.5	13.6	25.6	32.4	47.4	14.2
T5$_{top.k}$	14.4	31.5	30.4	47.8	10.6	26.4	29.9	47.7	13.7
T5$_{top.p}$	17.9	31.8	34.8	47	14.5	25.1	34.3	46.7	14.9
GPT2$_{gdy}$	18.2	49.9	38.7	59	19.2	39.4	38.5	58.8	41.3
GPT2$_{beam}$	18.9	48.4	38	58.6	18.8	38.8	37.8	58.4	42.7
GPT2$_{smp}$	13.5	30.4	30.7	46.2	10.4	24.5	30.6	45.9	14
GPT2$_{top.k}$	13.2	28.9	28.8	44.5	9.6	24	28.3	44.2	14.9
GPT2$_{top.p}$	14.1	29.7	30	44.7	10.9	22.9	29.9	44.5	15.4
BART$_{gdy}$	23.1	**51.1**	43.6	**61.4**	23.3	**42**	43.2	**61.3**	35.6
BART$_{beam}$	**24.6**	50.2	**44.2**	61.2	**24.3**	41.4	**44**	61	35.5
BART$_{smp}$	17	34.2	33.5	50	13.4	28	33	49.8	14
BART$_{top.k}$	18.4	33.4	35.8	49.6	15.2	27.2	35.5	49.3	16.1
BART$_{top.p}$	19.5	37.8	36.1	52.5	16.8	31.1	35.7	52.3	16.9
Input/Output Configuration 2									
T5$_{gdy}$	20.5	49.3	40.6	59.1	20.8	39.6	40.3	58.9	42.4
T5$_{beam}$	21.7	49.3	40.7	60	21.4	40.1	40.5	59.6	40
T5$_{smp}$	9.6	16.5	28.6	41.3	11.3	21.2	28.3	41.1	11
T5$_{top.k}$	8.3	16.9	27.3	41.3	9.2	21.4	27	41.1	12.3
T5$_{top.p}$	8.6	15.8	28.4	41.2	9.8	21.4	28.1	41	12.8
GPT2$_{gdy}$	17	49.8	37.8	59	17.1	39.2	37.5	58.8	43.4
GPT2$_{beam}$	19.9	49.6	39.4	59.2	20.4	39.5	39	58.9	40.9
GPT2$_{smp}$	14.9	28.5	31.3	44.3	11.3	22.7	30.9	44.2	13.1
GPT2$_{top.k}$	12.4	28.5	28.1	44	8.8	23.2	27.8	43.7	15.1
GPT2$_{top.p}$	12.9	29.4	30.2	44.8	9.4	23.3	29.9	44.7	14
BART$_{gdy}$	23.6	**51.4**	**44.7**	**61.5**	23.9	**42.3**	**44.4**	**61.4**	37.1
BART$_{beam}$	**24.5**	50.8	44.1	61.3	**24.3**	41.1	43.9	61.2	35.3
BART$_{smp}$	17.7	35.2	35	50.9	14.3	28.7	34.9	50.6	14.2
BART$_{top.k}$	16.5	33.4	34.2	49.6	13.9	27.7	34	49.4	14.1
BART$_{top.p}$	18.7	35.3	36.1	50.9	15.2	29.4	35.8	50.7	15.8
Input/Output Configuration 3									
T5$_{gdy}$	22.1	50.4	43	60.4	22.1	40.1	42.6	60.2	41.8
T5$_{beam}$	23.1	49.8	43.1	60.6	23.3	39.9	42.8	60.4	36.2
T5$_{smp}$	18.3	34.2	36.8	50.6	15.5	28.8	36.6	50.3	12.4
T5$_{top.k}$	17.6	31.9	35.2	48.5	14.3	26.2	35	48.3	14.1
T5$_{top.p}$	18.2	34.4	37.4	49.9	15.5	27.7	37.2	49.7	15.7
GPT2$_{gdy}$	20.9	51.3	43	60.7	22.4	40.7	42.9	60.5	42.2
GPT2$_{beam}$	22.1	51.8	43.2	61.2	22.7	41.7	43.2	61	42
GPT2$_{smp}$	17	33.8	35.5	49.3	12.8	27.9	35.4	49	14.2
GPT2$_{top.k}$	16.5	30.2	33.7	47	12.5	24.1	33.4	46.7	16.3
GPT2$_{top.p}$	15.5	33.4	33.4	48.4	11.4	27.1	33.2	48.3	15.9
BART$_{gdy}$	24.6	**52.9**	46.2	62.8	25.1	43.5	46	62.6	36.6
BART$_{beam}$	**26.9**	52.6	**47**	**63.5**	**27.1**	**43.9**	**46.9**	**63.3**	35.8
BART$_{smp}$	19.7	37.2	38	52.8	16.8	30.4	37.7	52.5	16.7
BART$_{top.k}$	18.5	33.6	36.2	50.2	14.8	27.9	35.7	50	15.8
BART$_{top.p}$	21.8	34.3	38.9	50.8	18.7	28	38.8	50.6	16.1

Table 2: Model results of automatic evaluation for generating IS, under the three options configurations with respect to BLEU, ROUGE F1 scores, and Repetition Rate (RR). Best Models are highlighted in gray for each option configuration.

tion 1 and 2. However, no major difference in RR was observed. The results suggest that providing target information helps generating more accurate implied statements.

Qualitative Evaluation. Finally, to get a better understanding of the quality and characteristics of the outputs, we manually analyzed a subset of ISs and TGs generated by the best-performing BART_{gdy} model (presented in Table 8 in Appendix). Examples (a) through (e) show that the model correctly distilled IS, despite differences in the word choices. The model also successfully predicted the TGs, although we occasionally observe a mismatch in both TGs and ISs (e.g., example (h)). This likely suggests that the model relies on strongly indicative keywords in HS (e.g., the word "*black*" in example (h)).

Wrongly generated TGs negatively impact the quality of the generated ISs, as shown in examples (h) and (i). Since TGs are generated first by the model during the generation, ISs generated next are highly constrained by those TGs. For instance, in example (i), the model first generated a wrong TG (WOMEN) and then produced an IS using the wrongly generated TG ("*women are HIV*"), instead of generating an IS about LGBT+.

We also observed that figurative form is a difficult phenomenon to handle. In fact, if an HS contains subtle sarcasm, IS tends to be generated incorrectly, e.g., examples (f) and (g). This pattern appeared throughout all three models' evaluations. In fact, detecting sarcasm is also a known challenge for the HS classification task (Justo et al., 2014; Frenda et al., 2022; Frenda, 2018; Badlani et al., 2019).

For the HS that does not explicitly state the TG or its describing words, the model struggles to generate the correct TG, e.g., in Example (i) the word "*gay*" is not mentioned. The possible cause of this drawback is the imbalanced TG distribution in the dataset. A similar observation is presented by Sap et al. (2019); the model can generate the correct IS if it has a high lexical overlap with the HS post, e.g., the examples (c) and (e).

6 Counter Narrative Selection

In this section, we present the task of finding appropriate counter-narratives for the hateful posts with the help of implied statements. Using IBSD and MT-CONAN datasets, we select a relevant CN for a given HS by utilizing the semantic similarities between elements of the two datasets.

6.1 Semantic Similarity Method

To compute the semantic similarity between textual elements from IBSD and MT-CONAN, we employed SentenceBert (Reimers and Gurevych, 2019), which provides sentence embeddings that can be compared using cosine similarity function. Our assumption is that if two HS posts are semantically close, then a CN paired with one HS can be appropriate also for the other [4].

6.2 CN Selection Strategies

To select proper counter narratives for the HS posts from IBSD, we experimented with various input texts.

$$\text{Sim}_{\text{base0}}=\text{SIM}(\text{HS}_{\text{IBSD}}, \text{HS}_{\text{MT-CON}} + \text{CN}_{\text{MT-CON}}) \quad (4)$$

$$\text{Sim}_{\text{base1}}=\text{FILTER}(\text{TG}_{\text{IBSD}}, \text{TG}_{\text{MT-CON}}) \mapsto \text{SIM}(\text{HS}_{\text{IBSD}}, \text{HS}_{\text{MT-CON}} + \text{CN}_{\text{MT-CON}}) \quad (5)$$

In the baseline configurations Equation 4 and Equation 5, the CN selection is conducted based only on HS, CN, and the target information, excluding the implied statement. SIM denotes a similarity function that takes two text inputs, converts them to numerical vectors, and outputs their cosine similarity score. In Equation 5, FILTER indicates a filtering function that keeps input pairs with the same target.

$\text{Sim}_{\text{base0}}$ (Equation 4) and $\text{Sim}_{\text{base1}}$ (Equation 5) configurations take two inputs to SIM function: an HS from IBSD (represented as HS_{IBSD}) and a set of concatenated texts of HS and CN pairs from MT-CONAN (represented as $\text{HS}_{\text{MT-CONAN}}+ \text{CN}_{\text{MT-CONAN}}$) as suggested by Chung et al. (2021b). SIM function then returns respective cosine similarity scores for all possible pairs between the two inputs. $\text{Sim}_{\text{base1}}$ configuration, additionally, applies target filtering before the similarity calculations. Using the BART_{gdy} model with Configuration 2 input from the IS generation task we grouped the data by TG attribute,

[4] For example, the HS from IBSD "*#Islam is designed to be a crushing of the human spirit with laws that only provide one outlet - the brutalizing of non Muslims.*" and the HS–CN pair from MT-CONAN "*Islam is a religion that promotes brutalities.*"—"*Islam is about love and compassion. I don't understand why you think otherwise.*" share a similar meaning, and thus the CN of the latter HS can be applied to the former.

where TGs are generated for IBSD while they are already provided for MT-CONAN. Each sample group with matching TGs are then inputted into SIM function. Performing the $\text{Sim}_{\text{base0}}$, i.e. the configuration without TG constraint, resulted in 188 examples of TG mismatch (micro F1 score of 84%). Instead, for $\text{Sim}_{\text{base1}}$, i.e. adding the TG constraint step before computing SIM, reduced more than half of the target mismatch (there are only 70 examples of TG mismatch deriving from a micro F1 score of 94% as can be seen in Table 7.). Since $\text{Sim}_{\text{base1}}$ outperformed $\text{Sim}_{\text{base0}}$ significantly, we will refer to it simply as Sim_{base} in the rest of the discussion. The remaining configurations take this distinction into account.

$$\text{Sim}_{\text{IS+base}}=\text{SIM}(\text{IS}_{\text{IBSD}} + \text{HS}_{\text{IBSD}}, \text{IS}_{\text{MT-CONAN}} + \text{HS}_{\text{MT-CONAN}} + \text{CN}_{\text{MT-CONAN}}) \quad (6)$$

$$\text{Sim}_{\text{base}}(\text{Filter}_{\text{IS}})=\text{FILTER}(\text{IS}_{\text{IBSD}}, \text{IS}_{\text{MT-CONAN}}) \mapsto \text{SIM}(\text{HS}_{\text{IBSD}}, \text{HS}_{\text{MT-CONAN}} + \text{CN}_{\text{MT-CONAN}}) \quad (7)$$

To explore the impact of using implied statements in identifying a proper counter narrative for a hateful post, $\text{Sim}_{\text{IS+base}}$ configuration (Equation 6) additionally prepends IS to SIM function inputs of Sim_{base} configuration. Although the majority of the input sequences has a length below the max sequence length of SentenceBert, i.e, 384, we chose inserting IS through prepending to make sure that it is included in the input when converted to the embeddings.

Finally, the last configuration, $\text{Sim}_{\text{base}}(\text{Filter}_{\text{IS}})$, (Equation 7) applies FILTER function, which keeps input pairs with the same implied statement. The FILTER function outputs at least five MT-CONAN samples for a given IBSD sample. To address the scenario when exact matching ISs of MT-CONAN are less than 5, we compute cosine similarity scores of the remaining samples and select the top $5 - x$ ISs to have the total of five candidates. Then, the Sim_{base} configuration is applied to the subset of samples derived from applying the IS-based filtering. Since the IS filtering step narrows down the search space of the CNs, we postulate that it would increase the probability of selecting a better CN for an HS post from IBSD.

6.3 Evaluation Metrics

Considering that there is no gold standard dataset to assess the quality of the selected counter narratives for the IBSD hate speech posts, we turned our attention to conducting a human evaluation. The set of samples to be evaluated are selected by (i) dropping samples with mismatched TGs, (ii) filtering out HSs from the IBSD with less than 50 characters, as short HSs are difficult to be interpreted by humans (e.g., "*Pedos going down*", "*Get this monkey off my back...*").

Finally, we run a human evaluation on 204 samples made by triplets, i.e., HS, CN_1 and CN_2.

Given a HS from IBSD, instead of letting human annotators to evaluate all three CNs selected by Equation 5,6, and 7, we present only two CN choices in a shuffled order. The annotators do not know which configuration the two CN choices come from, which prevents the annotator bias of detecting which configuration method the CN is selected from and enforce them to choose CNs purely based on the semantics.

Four human annotators, who have a research experience in hate speech and counter narratives, were given HS posts along with the associated TG information. Annotators were one male and three females, in terms of education level from PhD students to researchers. Two authors of the present work were involved in the task, the task was designed so to eliminate any possible confirmation bias (blind evaluation and randomization of the stimulus material). We also applied an adapted version of the guidelines by Vidgen et al. (2019) to safeguard the annotators' well-being against the risk of harmful consequences of working with abusive content.

Annotators had to select from two available CN choices, deciding which one better addressed the given HS. In case of a difficulty on understanding the given post, annotators could mark it as Not Applicable—'N/A'.

We have employed the **VictoryPointsTie** metric for the analysis of the human evaluation. This metric scores the annotators' CN choices as follows:

- **Victory**: if all annotators agree on the same CN, the selected CN configuration get assigned two points.
- **MajorityVictory**: if the majority agrees on the same CN, then its configuration get assigned two points. Majority is defined as 3

out of 4 annotators agree on the same CN or 2 annotators agree on the same CN while others vote for 2 separate choices (including 'N/A' option).

- **Tie**: if 2 annotators vote for the same CN and the other 2 vote together for another CN (including 'N/A' option), then each configuration counts as one point.

To assess if a particular configuration rises to prominence, the total points are summed for each configuration using the above metric. In the case of the same CN selected by 2 different configurations (ensemble configurations), we also assign victory points to the corresponding ensemble configuration. Through the ensemble configuration analysis, we aimed at understanding if jointly decided CN selections are more reliable than CN selections done by a single configuration or certain ensembles are more reliable than others on finding a better CN for an HS post. We used Cohen's kappa (Cohen, 1960) for the inter-annotator agreement.

7 Results

The results of the human evaluation are shown in Tables 3 and 4. $\text{Sim}_{\text{IS+base}}$ outperforms other configurations, while $\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$ performs poorly. The best configuration includes all three elements of a data sample, i.e., IS, HS, and CN, with IS weakly constraining the produced embedding. Instead, the IS filtering in $\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$ configuration acts as a hard constraint, resulting in a limited number of CNs to be fed to the SIM function. This configuration assumes that constraining by IS narrows down the CN set to keep, and these would likely be the most relevant ones. However, human evaluation results show that keeping only the top five most similar ISs in this configuration limits the CN search space too much. Even if there were many ISs exactly matched between IBSD and MT-CONAN, the FILTERING function would discard all the other highly similar ISs and thus some possible good CNs.

Among the samples with one of two CN options chosen by an ensemble configuration (total of 99 samples), the CN supported by an ensemble is preferred (68.7%) over the other CN that was selected by only one configuration (31.3%).

The most reliable ensemble configuration was ($\text{Sim}_{\text{base}} \cap \text{Sim}_{\text{IS+base}}$). As shown in Table 5, it was chosen by comparing the percentages of the counts

Configuration	Score
Sim_{base}	187
$\text{Sim}_{\text{IS+base}}$	**203**
$\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$	161

Table 3: Computed VictoryPointsTie score for each configuration.

Configuration pair	Config	#
($\text{Sim}_{\text{base}} \cap \text{Sim}_{\text{IS+base}}$, ($\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$))	Ens.	25
	Sing.	8
($\text{Sim}_{\text{IS+base}} \cap (\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}})$, Sim_{base})	Ens.	22
	Sing.	10
($\text{Sim}_{\text{base}} \cap (\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}})$, $\text{Sim}_{\text{IS+base}}$)	Ens.	21
	Sing.	13
Ensemble configs total		**68**
Single configs total		31
Total		99

Table 4: The counts of selected samples for each configuration pair from the human evaluation dataset samples with at least one CN choice selected by two configurations. Configuration pair consists of (Ensemble Config, Single Config). **Ensemble Config** refers to the configuration combination where two configurations selected the same CN . **Single Config** indicates the remaining configuration that did not have any overlap of selected CN with other configurations.

of the selected CNs from "Ensemble Config" in the total counts of CNs for the particular configuration pairs ("Ensemble Config" + "Single Config"). This observation aligns with the earlier finding that the $\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$ configuration performs the worst.

Ensemble configuration	%
$\text{Sim}_{\text{base}} \cap \text{Sim}_{\text{IS+base}}$	**75.8**
$\text{Sim}_{\text{IS+base}} \cap (\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}})$	68.8
$\text{Sim}_{\text{base}} \cap (\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}})$	61.8

Table 5: The percentages of each ensemble configuration count out of the total configuration pair computed using Table 4.

The annotations in the human evaluation dataset had the majority agreement of 85.6% and perfect agreement of 39.6%, which demonstrate a fair inter-annotator agreement with a Cohen's Kappa $= 0.4$.

To sum up, the results show that applying the TG filtering drastically deteriorates the CN selection performances. Using the implied statement helps choosing relevant counter narratives, despite

coming with some caveats. Incorporating the IS, while preserving the original content (HS/CN) in the input (such as $\text{Sim}_{\text{IS+base}}$), allows the embedding to focus more on its semantic contents. However, if the IS is used as a hard filtering (such as $\text{Filter}_{\text{IS}}+\text{Sim}_{\text{base}}$), it could overly confines the available CN candidate space.

8 Conclusion

In this paper we present a novel approach for selecting counter narratives to fight hate speech based on the use of the HS implied statements. First, we distill the implied statement from the hate speech through a generation task with testing several LMs/decoding methods.

Concerning the IS generation results, the fine-tuned BART model with the greedy decoding method yields the best BLEU and ROUGE scores. Then, to retrieve possible counter narratives for hate speech posts, we compute semantic similarity between two input texts (with varying formats including HS, CN, hate target, and implied statements), given the assumption that semantic similarity and human judgment correlate. The human evaluation results show that filtering by TG reduces the mismatch between HS and their selected CN. While the CNs selected from the configuration that included IS as a part of the similarity calculation was preferred, filtering by IS constrains too many utilizable CN candidates. In addition, a joint selection of the same CN by two different configurations (ensemble) yielded even better results in terms of human preference.

Although the novel approach of incorporating IS and TG show effectiveness in the CN selection task, there are several possible directions that can be explored. Addressing target imbalance in the dataset and out-of-target implied statement generation for evaluating the model performances on zero-shot out of domain experiments could be explored in the next steps. Additionally, more complex semantic similarity methods could be integrated into the task.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9:147–169.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6).

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th international conference on World Wide Web companion*, pages 759–760.

Rohan Badlani, Nishit Asnani, and Manan Rai. 2019. An ensemble of humour, sarcasm, and hate speech-for sentiment classification in online reviews. In *EMNLP*.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention.

Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. *ArXiv*, abs/2010.12472.

Mauro Cettolo, Nicola Bertoldi, and Marcello Federico. 2014. 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*, pages 166–179, Vancouver, Canada. Association for Machine Translation in the Americas.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021a. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021b. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. *arXiv preprint arXiv:2109.05322*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *ACL*.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroglu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51:1 – 30.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Simona Frenda. 2018. The role of sarcasm in hate speech.a multilingual perspective.

Simona Frenda, Alessandra Teresa Cignarella, Valerio Basile, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing explicit and implicit hate speech using a weakly supervised two-path bootstrapping approach. *arXiv preprint arXiv:1710.07394*.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint arXiv:2106.00742*.

Raquel Justo, Thomas Corcoran, Stephanie M Lukin, Marilyn Walker, and M Inés Torres. 2014. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Y. Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, A Azatian, Aadila Hussain, Alejandro Lara, olmos g, Asmaa Al Omary, Christina Park, C. C. Wang, X Wang, Y. Zhang, and Morteza Dehghani. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jitendra Malik, Guansong Pang, and Anton van den Hengel. 2022a. Deep learning for hate speech detection: A comparative study. *ArXiv*, abs/2202.09517.

Jitendra Singh Malik, Guansong Pang, and Anton van den Hengel. 2022b. Deep learning for hate speech detection: A comparative study. *arXiv preprint arXiv:2202.09517*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

Chikashi Nobata, Joel R. Tetreault, Achint Oommen Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Jing Qian, Mai ElSherief, Elizabeth M. Belding-Royer, and William Yang Wang. 2018. Hierarchical cvae for fine-grained hate speech classification. In *EMNLP*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings*

of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Tanya Silverman, Christopher J Stewart, Zahed Amanullah, and Jonathan Birdwell. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*, 54.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*.

Serra Sinem Tekiroglu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Matthew Williams. 2019. Hatred behind the screens: A report on the rise of online hate speech.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the first workshop on abusive language online*, pages 57–62.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7.

Wenjie Yin and Arkaitz Zubiaga. 2022. Hidden behind the obvious: misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625*.

Marc Ziegele, Pablo Jost, Marike Bormann, and Dominique Heinbach. 2018. Journalistic counter-voices in comment sections: Patterns, determinants, and potential consequences of interactive moderation of uncivil user comments. *SCM Studies in Communication and Media*, 7(4):525–554.

Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis. *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*.

A Appendix A - Data Filtering details

This section explains in detail the data filtering and pre-processing techniques we applied to the datasets that are described. The original datasets contains $153,498$ samples from IHC and SBIC for IS generation task and $5,003$ samples from MT-CONAN for CN selection task.

Pre-Filtering Steps. Data preparation required several steps. (i) As the first step we selected only sentences marked as "implicit hate speech" from IHC (since only this category was annotated with the IS). For SBIC, we kept examples annotated as intentional and offensive. (ii) Then we selected samples with single-sentenced ISs. (iii) As a third step we chose those samples that have a target that aligns with targets existing in MT-CONAN dataset[5]. We also discarded inter-sectional examples (i.e. targeting two or more groups simultaneously, such as POC and WOMEN in "*Black women are only able to cook fried chicken.*") (iv) We then checked for possible duplicates, however we allowed the same HS post to have more than one ISs. After these filtering steps, we obtained $35,923$ samples of unique HS–IS pairs.

Standardization Steps. Although both (Sap et al., 2019) and (ElSherief et al., 2021) claim to strictly follow Hearst's pattern for annotation, MTURK workers did not always follow the instructions. For this reason we applied a fifth filtering/standardization step to obtain examples with an IS that rigorously follow the pattern $\langle subject \rangle$ – $\langle predicate \rangle$ – $\langle object \rangle$, with *object* being optional. Additionally, it is important that IS contains TG as the subject. To check for this, we utilized Stanza library[6] from Stanford to perform dependency parsing, which returns the grammatical relationships within the sentence. We checked whether the sentences contained a nominal subject (*nsubj*) and if so, whether they matched the originally annotated TG names. We considered a match if any of the following conditions is true: the exact word matches with the annotated TG name, or the extracted subject is one of the collected terms used to represent TG. For instance, if the TG is MUSLIM, then we consider a match if the annotated IS has the subject "*muslim*", "*islamic people*", or "*muslim people*." We further applied lemmatization and stemming to make sure the different forms were caught, e.g., *illegals* → *illegal*, *woman* → *women*.

We also kept IS whenever its first word matched the TG, but the parser did not recognize it as a noun subject of the sentence. Instead we discarded IS when the TGs was present in the sentence but not marked as subject, e.g., "*people hate **Jews***", "*America needs to control the **black population***."

For those examples in which TGs were not detected in the ISs, we manually reviewed the sentence structure. For instance, we checked part of speech using the Stanza tool to see if we could supply TG into the IS. The following rules were used to decide about keeping the examples:

- Some examples had a sentence starting with a "*group*" noun (e.g., "*Group does drugs*", "*Group is worthless*"). We simply replaced that word with original TGs, while fixing the changed IS to correct verb/auxiliary verb (e.g., "*Black people do drugs*", "*Women are worthless*").

- If the first word was a pronoun (e.g., "*Their lives do not matter*", "*They are property of men*"), then we also replaced it by the TG (e.g., "*Black lives do not matter*", "*Women are property of men*").

- In case the first word was an auxiliary verb, (e.g., "*Are all terrorists*", "*Are just objects*"), then we inserted the original TG at the beginning of sentence (e.g., "*Muslims are all terrorists*", "*Women are just objects*").

- If the IS was a one-word adjective (e.g., "*Feminine*", "*Stupid*", "*Lazy*") then we prepended TG followed by "*are*" to it to build a complete sentence (e.g., "*Gays are feminine*", "*Jews are stupid*", "*Black people are lazy*").

- If the beginning of the sentence started with an adverb and the second word was a verb or adjective (e.g., "*Often harmed*", "*Always fail exams*", "*Easily offended*") then we appended TG at the beginning of a sentence. We needed to also append "are" since some take the form of past tense VERB (e.g., "*Muslimes are often harmed*", "*Blacks always fail exams*", "*Women are easily offended*").

[5]These targets are {WOMEN, POC, JEWISH, MUSLIM, MIGRANT, LGBT+, DISABLED}. Since the original SBIC dataset did not use unique or standardized labels (e.g. "*dark-skinned men*", "*blacks*","*black folks*") we created a mapping to our standardized labels. See Table 6 for some of the example words representing the considered TGs.

[6]stanfordnlp.github.io/stanza

TG	example words
POC	black folks, africans, people of color, african
LGBT+	gay men, lesbian women, trans men, gay people
JEWISH	jewish folks, jews, jewish people, jewish
MUSLIM	muslim folks, islamic folks, muslims, islam
WOMEN	feminists, woman, women
MIGRANTS	immigrants, refugees, illegal immigrants, illegal aliens, immigrant
DISABLED	mentally disabled folks, physically disabled folks, blind people, folks with down syndrome

Table 6: Sample words that were commonly used in the SBIC and IHC dataset to refer each TG labels

- If the second word of the sentence was adjective, adposition, or adverb (e.g., "*Only good at sports*", "*Overly religious*") then we prepended TG followed by "are" at the beginning of the sentence (e.g., "*Black people are only good at sports*", "*Muslims are overly religious*").
- Some sentences start with the particle "not" (e.g., "*Not attractive*", "*Not intelligent*", "*Not get along with others*"). In this case, we fixed the ISs by appending TG and "don't" if the second word was a verb, otherwise appending "are" (e.g., "*Gays are not attractive*", "*Black people are not intelligent*", "*Women don't get along with others*").

B Appendix - Training Details

All models were trained for 5 epochs. The training objective was to maximize the sum of the metrics at the evaluation time with a preference for a higher score. A batch size of 8 was selected for training and 16 for inference. The gradient accumulation step allows for accumulating gradients and performing the model's optimization step afterward. We used a step size of 16 to increase the overall batch size. The default learning rate of 2e-05 was used for GPT-2, whereas 5e-05 for BART and T5. We framed the IS generation as a summarization task since it was one of the original pre-training tasks of T5 and fine-tuned the model with a *"Summarize:"* source prefix. This decision was taken after we conducted a small experiment by fine-tuning a small subset of the IBSD dataset with and without a source prefix, as well as a custom prefix ("*explain implied statement:*"). The results showed that having a prefix is beneficial for small data (in this experiment 1,000 samples); however, the choice of which prefix to use did not make a major difference in their performance.

At inference time we used various decoding methods: Greedy search, Beam search (beam width was set to 3), Temperature sampling with a temperature value of 0.9 (Ackley et al., 1985), Top-k sampling, with k value was set to 40 (Fan et al., 2018), and finally Top-p sampling with p value equal to 0.9 (Holtzman et al., 2020).

Models	DISAB.	POC	WOMEN	JEWISH	MIGR.	MUSLIM	LGBT+	MacroF1	MicroF1
T5$_{gdy}$	80	93.8	93.5	93.9	85.4	91.4	90.1	89.7	92.2
T5$_{beam}$	81.5	95.1	92.9	95.2	85.9	93.2	88.9	90.4	92.9
T5$_{smp}$	49.1	54.2	45.5	48	40.3	50	58	43.1	34.3
T5$_{top.k}$	29.1	50.6	52.2	55.3	46.3	49.7	40.6	40.5	34.7
T5$_{top.p}$	35.3	48.6	44.4	50	39.7	43.7	44.4	38.3	30.6
GPT2$_{gdy}$	83.5	92.4	91.6	93.4	76.1	91.8	79.5	86.9	90.3
GPT2$_{beam}$	81.4	93.2	90.4	93.1	78.8	91.1	82.4	76.3	90.2
GPT2$_{smp}$	74.1	91.7	89.1	92.3	74.4	89.5	73.5	73.1	88
GPT2$_{top.k}$	74.7	91.3	89.2	89.8	72.3	89	71.1	72.2	87.4
GPT2$_{top.p}$	75	92	89.5	92.9	77.8	90.4	76.4	84.9	89
BART$_{gdy}$	**88.6**	**95.4**	94.9	95.2	**89.2**	95.5	90.5	**92.8**	**94.3**
BART$_{beam}$	86.4	95.1	94	**95.5**	88.3	95.1	91.5	92.3	93.9
BART$_{smp}$	87.2	94.9	94.5	95.3	87.3	94.6	**93.6**	92.5	93.9
BART$_{top.k}$	**88.6**	94.7	94	95	86.2	94.2	89.6	91.8	93.4
BART$_{top.p}$	87.5	95.4	**95.1**	95	89	**96**	88.7	92.4	94.2

Table 7: Model results of automatic evaluation for generating TG given HS (Option 2).

	HS	predicted TG	predicted IS	referenced TG	referenced IS
(a)	If only Africa had more mosquito nets Then we could save millions of mosquitoes every year from dying of AIDS	POC	africans are all diseased	POC	"black folks carry diseases", "africans deserve to die first"
(b)	Allah is probably the true god Because the universe started with an explosion	MUSLIM	muslims are terrorists	MUSLIM	"muslims are terrorists", "the muslim religion is violent", "muslims are barbaric"
(c)	no such thing as might they ““must be deported.. all illegal aliens must be deported.no amnesty, or we will face unprecedented negative consequences.	MIGRANTS	immigrants should be deported	MIGRANTS	"all illegal immigrants must be deported"
(d)	Why do the handicapped make poor lawyers? Their cases don't have a leg to stand on	DISABLED	physically disabled folks are useless	DISABLED	"disabled people are incapable"
(e)	There is nothing I would love to see more than the arrest, trial and execution of these murderous and genocidal Zionists!	JEWISH	jewish people are murderers	JEWISH	"jewish people are violent and genocidal", "jewish people are evil", "jewish people are murderous"
(f)	I'm going to name my next cat äutistic"...because that's actually funny.	DISABLED	autistic people are funny	DISABLED	"mentally disabled folks laughed at", "autistic people are dumb", "autistic people are often teased", "autistic people are easy to make fun of", "mentally disabled folks doesnt deserve empathy"
(g)	Anyone else find it sexist that trans women can't have children?	LGBT	trans women are not equal to men	LGBT	trans women aren't real women, trans women are fake
(h)	I have blue eyes. I got them from my dad. My mum has black eyes. She also got them from my dad.	POC	black folks are criminals	WOMEN	"women are not equal", "women should be disciplined by they're husbands", "women are subservient to men "
(i)	jack probably has hiv which is why he's doing this because he's jealous that we all don't have hiv because we don't fuck people in the butts	WOMEN	women are hiv	LGBT+	"gay people have diseases", "gay men have hiv", "gay people envy straights", "gay men are responsible for hiv"

Table 8: Examples generated by the BART$_{gdy}$ generation model. The model struggles to understand sarcasm (g) and directly reuse words in HS(f), but it is possible to infer the implications of the HS(a, b, d).