

# Consolidating Strategies for Countering Hate Speech Using Persuasive Dialogues

Sougata Saha and Rohini Srihari

State University of New York at Buffalo

Department of Computer Science and Engineering

{sougatas, rohini}@buffalo.edu

## Abstract

Hateful comments are prevalent on social media platforms. Although tools for automatically detecting, flagging, and blocking such false, offensive, and harmful content online have lately matured, such reactive and brute force methods alone provide short-term and superficial remedies while the perpetrators persist. With the public availability of large language models which can generate articulate synthetic and engaging content at scale, there are concerns about the rapid growth of dissemination of such malicious content on the web. There is now a need to focus on deeper, long-term solutions that involve engaging with the human perpetrator behind the source of the content to change their viewpoint or at least bring down the rhetoric using persuasive means. To do that, we propose defining and experimenting with controllable strategies for generating counter-arguments to hateful comments in online conversations. We experiment with controlling response generation using features based on (i) argument structure and reasoning-based Walton argument schemes, (ii) counter-argument speech acts, and (iii) human characteristics-based qualities such as Big-5 personality traits and human values. Using automatic and human evaluations, we determine the best combination of features that generate fluent, argumentative, and logically sound arguments for countering hate. We further share the developed computational models for automatically annotating text with such features, and a silver-standard annotated version of an existing hate speech dialog corpora.<sup>1</sup>

## 1 Introduction

Hateful comments are prevalent on social media platforms (Schmidt and Wiegand, 2017) and of-

ten go unchecked. While most available solutions to tackling online hate resort to brute force methods such as content flagging and blocking, by not addressing the underlying cause, such methods usually provide superficial and temporary respite. They often yield contradictory outcomes by further agitating the perpetrator, thus fueling hate, and even pose as restraints to online free speech, which leads to demonstrations. The steep growth of online hate speech is difficult for human regulators to keep up with, which is further aggravated by public access to large language models (LLMs) (Zhao et al., 2023) that are capable of generating fluent yet synthetic content at scale. Hence, there is a need for deep and scalable means for amicably addressing hate speech. Instead of only being reactive, a solution should aim at changing the perception of the initiator of hate speech. It should try to minimize the rhetoric by either persuading the perpetrator of hate speech contrary to their standpoint or at the least trying to broaden their perception and making them aware of the other facets of the topic of discussion. Online conversational agents can be such a robust solution to countering online hate.

Persuasive conversations that strive to change the mental state of an interlocutor rely on subtler human-based appeals. According to the ancient Greek philosopher Aristotle, persuading an audience depends on how well a person appeals in three areas: logos, ethos, and pathos (Rapp, 2002). Recent research has engendered general-purpose LLMs like ChatGPT (Ouyang et al., 2022), GPT-4, Bard (Thoppilan et al., 2022), and inter-alia, whose fluent conversational capabilities have left the public in awe. Although impressive, without being explicitly trained to persuade, their efficacy in the purposeful and nuanced human task of persuasion is questionable. Hence, there is a need to develop controllable conversational systems that can incorporate such appeals while generating persuasive counter-hate responses. Here we attempt

<sup>1</sup>OFFENSIVE CONTENT WARNING: This report contains some examples of hateful content, which could be offensive and cause you distress. It is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please contact the authors for code.

to consolidate and establish a set of features from diverse sources that can aid control and guide computational models to construct counter-arguments to hateful comments in online conversations. We broadly identify two feature categories for capturing the different aspects of a response: (i) Value-based: Features based on human qualities and characteristics such as Big-5 personality traits (Soto, 2018; Costa Jr, 1992) and human values (Rokeach, 1973; Schwartz, 1994). (ii) Structure-based: Features based on the patterns of reasoning such as Walton argument schemes (Walton et al., 2008) and argumentative discourse-based speech acts (Stolcke et al., 2000) such as denouncing, questioning, etc. We incorporate a two-step process where first, we automatically annotate available hate speech dialogue datasets with the value and structure-based features using computational models. Second, using the annotated features as guides, we experiment with computational models for generating counter-arguments to hate speech. We experiment with different feature combinations to determine the best features for guiding counter-argument generation and report automatic and human evaluation results. In the process, we yield a silver-standard annotated version of the hate speech dialog corpora, with additional computational models for automatically annotating text with the defined features. We summarize our contributions as follows:

- We derive features from psychology, linguistics, philosophy, and social science and guide computational models for counter-hate argument generation.
- We release a computational framework for automatically identifying personality traits, human values, argument schemes, and argument types from argument text.
- We experiment and analyze the best combination of the features for response generation.

Please note that this paper is strictly linguistically motivated and does not compose any psychographic or demographically motivated data profiling, analysis, or segmentation. The proposed methods only utilize diverse sources to derive features based on single turns of short text around specific topics, with the motivation being identifying observable stylometric differences between hateful and counter-hate response text. We use the features as control codes (Keskar et al., 2019) for

aiding language modeling for counter-argument generation and hypothesize that such control codes can better partition the learned probability distribution. Furthermore, since the datasets used in the experiments are created synthetically by experts, the derived textual features do not portray the actual traits of users in a live setting. Through our experiments, we only highlight the differences in the presence of features, and the actual labels do not matter for our purpose.

## 2 Related Work

There has been noticeable advancement in gearing technology to curb the menace of online hate (Kiritchenko et al., 2021; Windisch et al., 2021). Zhu and Bhat (2021) proposed a three-module pipeline approach where they generate multiple counter-arguments, and filters using a BERT-based model (Devlin et al., 2018), and select the most relevant response using a novel retrieval-based method. Saha et al. (2022a) implemented CounterGeDi, an ensemble approach to enable DialoGPT (Zhang et al., 2019) generate polite, detoxified, and emotionally laden counterspeech. de los Riscos and D’Haro (2021) introduced Toxicbot, a system that detects and generates responses to intervene during online conversations with hate speech content. (Chung et al., 2021b) introduced an external knowledge-based counter-narrative generation pipeline that can generate factual arguments across multiple domains. Chung et al. (2020) leveraged GePpeTto for effectively generating counter-narratives in Italian.

Most available counter-hate speech corpora are limited to single-turn pairs of hate speech and counter-hate response, while very few focus on multi-turn dialogues. Tekiroglu et al. (2020) presented a study on using GPT-2 (Radford et al., 2019) for generating responses to hate and produced a silver-standard annotated dataset. Chung et al. (2019) introduced Conan, which contains hate speech/counter-narrative pairs in English, French, and Italian about Islamophobia. Using a Human-in-the-Loop approach, Fanton et al. (2021) released Multi-Target Conan. They expanded Conan to include English examples across multiple hate targets. Chung et al. (2021c) released a knowledge-grounded subset of Multi-Target Conan examples. Furman et al. (2022) collected responses to the hateful comments in the Hateval corpus (Basile et al., 2019) and further enriched it with annotations based on Wagemann’s periodic table of argu-

ments (Wagemans, 2016). Ashida and Komachi (2022) employed LLMs to create Chasm, a synthetic and quality-controlled counter-hate dataset. They provided further annotations that can also facilitate research in counter-narrative evaluation. Bonaldi et al. (2022) shared DialoConan, a multi-turn synthetically generated and quality-controlled dialogue dataset between a hater and an NGO operator. We use DialoConan for our experiments and incorporate diverse features as control codes (Keskar et al., 2019) to better guide counter-argument generation.

### 3 Defining and Annotating Features

As described in Table 1, we derive features from psychological and characteristic-based Big-5 personality traits (Sanford, 1942; Saha et al., 2022b) and human values (Schwartz, 1994; Kiesel et al., 2022), logic and reasoning-based argument schemes (Walton et al., 2008; Kondo et al., 2021) from philosophy, and linguistic speech act-based (Stolcke et al., 2000) counter-argument types (Chung et al., 2019) and bucket them into two broad categories: Value and Structure, for our purpose. Overall we use 20 distinct features in our experiments and automatically annotate DialoConan with such features using computational models, which we discuss below.

#### 3.1 Big 5 Personality Traits

Personality is the most fundamental dimension of variation between humans, and it greatly influences our situational reactions, thoughts, feelings, expressions, and actions (Mairesse et al., 2007). We use the Big-5 personality trait (Soto, 2018; Costa Jr, 1992) classifier released by Saha et al. (2022b) to identify the most likely personality traits of the proponent of a hate speech text. We use the parser to parse each turn independently and annotate the hate and counter-hate argument with the Big-5 traits. Figure 3 (Appendix A) illustrates the distribution of each trait by the type of speech. We observe hateful comments to be classified as more neurotic, whereas counter-hate responses contain higher proportions of the other four classes.

#### 3.2 Human Values

Human values (Rokeach, 1973; Schwartz, 1994) are a set of beliefs that condition human being and influences their actions in different scenarios. It is a system based on cultural, social, and personal fac-

tors and varies across persons. Although they are well-researched in social science and already studied in formal argumentation settings (Bench-Capon, 2003), it is only recently that computational models (Kiesel et al., 2022) have been used to identify them from arguments. Kiesel et al. (2022) defined 20 human value categories (L2 categories) that capture four broad aspects (L3 categories) of values: openness to change, self-enhancement, conservation, and self-transcendence. Each L2 category encapsulates finer distinct human values (L1 category) that better characterizes each L2 category. For example, the L2 category of ‘Self-direction: thought’ represents the values ‘Be creative’, ‘Be curious’ and ‘Have freedom of thought’. Furthermore, each value is described by a set of sentences (value descriptors, 218 in total) explaining what it means to possess such a human value. For example, ‘promoting imagination’ and ‘being more creative’ are two of the several descriptors for the value ‘Be creative’. We train computational models to automatically identify the L2 values from the hateful and counter-hateful text and use them as value-based features. We use the standard training and validation splits from the SemEval shared task by Mirzakhmedova et al. (2023) and include the samples from the validation-zhihu split in the training samples, which are derived from the Chinese question-answering website Zhihu<sup>2</sup>. Due to class imbalances, we limit ourselves to the top 6 value categories: Achievement, Benevolence: caring, Security: personal, Security: societal, Self-direction: action, and Universalism: concern out of the twenty categories defined by Matt. Inspired by Saha and Srihari (2023b), we implement a majority-based ensemble approach comprising three models, which we discuss below.

##### 3.2.1 Classification Based Model

We fine-tune Roberta large (Liu et al., 2019) with a multi-task objective of predicting L1, L2, and L3 labels from arguments. We encode each example using Roberta’s pooler representation, and for each level, we use a linear layer to yield the logits. The model is optimized end-to-end by minimizing a weighted averaged Binary Cross-Entropy (BCE) loss function, with 0.23, 0.33, and 0.44 as the respective weights. We use AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 1e-5, and train until the validation loss stops decreasing

<sup>2</sup><https://www.zhihu.com/explore>

Type	Source	Derived From	Features / Control Codes
Value	Big-5 personality	Psychology & Social Science	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
	Human Values	Psychology & Social Science	Achievement, Benevolence: caring, Security: personal, Security: societal, Self-direction: action, Universalism: concern
Structure	Argument Scheme	Philosophy & Linguistics	From consequence, From source authority/knowledge, Goal from means/means for goal, Rule or principle
	Argument Type	Linguistics Speech Acts	Denouncing, Facts, Hypocrisy, Positive, Question

Table 1: Description of different types of features.

for 4 consecutive epochs. We use the standard training and validation splits and additionally merge the validation-zhihu split in the training samples, yielding 5,493 training and 1,896 validation examples.

### 3.2.2 Entailment Based Model

We implement an entailment-based model for identifying the L1 value descriptors that an argument text entails. We transform the dataset conducive for textual entailment by creating positive and negative argument and value descriptor pairs and aggregate the predictions at the L2 level. We follow the same approach as [Saha and Srihari \(2023b\)](#) to create the entailment pairs. We also merge the validation-zhihu split with the training samples, resulting in 189,312 training and 65,900 validation examples. The entailment model comprises a Roberta-base encoder and a linear layer to predict the logit. During training, we concatenate an argument and value descriptor text and pass the encoder pooled representation through a linear layer to predict the logit. Similar to [Saha and Srihari \(2023b\)](#), we load the model weights with pre-trained MNLI weights and fine-tune by minimizing the BCE loss function using AdamW optimizer and a learning rate of 1e-5.

### 3.2.3 Similarity Based Model

We implement an embedding similarity-based approach for determining which of the 218 value descriptors an argument text most resonates with. First, we train a metric learning-based model to learn embeddings of the 218 L1 value descriptors. Next, treating the descriptor embeddings as fixed centroids, we train an embedder that transforms and embeds argument text to the embedding space of the centroids. Finally, we determine the best descriptor of an argument text by computing the cosine similarity between its embedding and the descriptor centroids and aggregate the predictions at the L2 level.

**Step 1: Generate human value descriptor embeddings:** We create multiple training quadruples per value descriptor, consisting of the descriptor as the anchor element, one positive, and two negative descriptors. The positive descriptor belongs to the same L1 class as the anchor. For the negative descriptors, we randomly sample one descriptor from the same L2 but different L1 class and another from a different L2 and L1 class altogether, which we term as hard and easy negatives. Overall we create 702 examples and randomly use 90% examples for training, and remaining for validation.

We learn embeddings using a Roberta-based encoder by optimizing a metric learning-based loss function. The training objective is two folds: (i) Minimizing the distance between the anchor and the positive descriptor embedding while maximizing the distance between the anchor and the easy-negative descriptor embedding, which enables descriptors of the same aspects (L1) of human values (L2) to be closer in the embedding space than dissimilar ones. (ii) Minimizing the distance between the hard-negative and anchor descriptor while maximizing the distance between the hard-negative and the positive descriptor embedding, which enables descriptors of the different aspects (L1) of the same human values (L2) to be close to the anchor while still being far enough in comparison to the descriptors of the same aspects in the embedding space. The below set of equations formally details the learning algorithm.

$$\text{enc}(x) = \text{Roberta\_base\_pooler}(x) \quad (1)$$

$$a = \text{enc}(\text{anchor}) ; p = \text{enc}(\text{positive}) \quad (2)$$

$$n_x = \text{enc}(x \text{ negative}), x \in \{\text{easy}, \text{hard}\} \quad (3)$$

$$D(x, y) = \text{cosine\_similarity}(x, y) \quad (4)$$

$$L = \alpha * [D(a, p) - D(a, n_{\text{easy}})] + \beta * [D(n_{\text{hard}}, a) - D(n_{\text{hard}}, p)] + M \quad (5)$$

Where M is a fixed margin that is set to 1.0, and



$\alpha=2.0$ ,  $\beta=1.0$ , and are chosen empirically. We optimize the model using AdamW optimizer with a learning rate of  $2e-5$ , and train until the validation loss stops decreasing for 5 consecutive epochs.

**Step 2: Train argument embedder:** We train an argument embedder for transforming and embedding an argument text to the embedding space of the value descriptors. For every example, we create multiple positive and negative argument-value descriptor pairs. If there are  $K$ -positive descriptors per argument, we ensure a balanced negative descriptor by randomly sampling one value descriptor per L1 class and persisting  $K$  random values. The training and validation dataset comprises 200,059 and 69,607 argument-value descriptor pairs. We pass the Roberta pooler representation of an argument text through 3 fully-connected layers with ReLU activation and train by minimizing the cosine distance between its embedding and positive value descriptor. We train the model end-to-end and use the same training parameters as the value descriptor embedder. During inference, we determine the best descriptor of an argument text by computing the cosine similarity between its embedding and the descriptor centroids and aggregate the predictions at the L2 level.

### 3.2.4 Results

Table 2 shares the results for each of the three models, along with a majority-based ensemble of the three models. We report the macro-averaged F1 score for the positive class and overall level for all the twenty and top 6 human value classes. Overall the ensemble approach almost always yields better results. The classification-based model has the best validation results of the three individual models, which we attribute to its more parameters. We use the ensemble for automatically annotating each turn’s hate and counter-hate text of the target dataset with the top 6 human values. Figure 4 (Appendix A) plots the distribution of each value. We observe hateful comments to exhibit significantly more societal security concerns, whereas counter-hate responses manifest higher personal security and care.

## 3.3 Argument Schemes

Argument schemes are typical reasoning and inference patterns found in arguments. Walton provided an in-depth study of argument schemes (Walton et al., 2008) and defined 60 such schemes prevalent in daily argument text. Kondo et al. (2021)

Model	All Classes		Top Classes	
	Positive	Macro	Positive	Macro
Classification	0.48	0.70	0.60	<b>0.73</b>
Entailment	0.47	0.66	0.61	0.68
Similarity	0.44	0.64	0.56	0.63
Ensemble	<b>0.52</b>	<b>0.71</b>	<b>0.63</b>	0.72

Table 2: F1 scores on validation set for Human Value Detection from text.

defined Bayesian networks to represent the inference pattern of arguments, which inspired Saha and Srihari (2023a) to experiment with computational models to detect the type of reasoning used by an argument text. They restricted to 6 of Walton’s argument schemes: “Means for Goal”, “Goal from Means”, “From Consequence”, “Source Knowledge”, “Source Authority”, and “Rule or Principle”. We use their computational model to identify the most likely scheme used by each turn’s hate and counter-hate argument. Based on the distribution of the labels, we further combine the classes “From Source Authority” with “From Source Knowledge” and “Goal for Means” with “Means for Goal”. Figure 5 (Appendix A) plots the distribution of each scheme. We observe that counter-hate arguments are significantly more authoritative and knowledge-based.

## 3.4 Argument Type

Identifying the speech act of an argument can be crucial and provide more information about what kind of arguments generally work in a hateful scenario. There are scenarios in toxic discussions where responding with a fact might be more effective in reducing the rhetoric than just denouncing and challenging the standpoint of the hateful argument. For example, the comment ‘immigrants are eating into the American job market and should be banned’ can be countered with a fact around how immigrants only take up those jobs that Americans don’t, instead of just denouncing it. Chung et al. (2019) identified and defined ten categories of counter-responses to Islamophobic hateful comments: affiliation, consequences, denouncing, facts, humor, hypocrisy, positive, question, negative, and others. Due to the skewed class balance, we limit to the first 6 categories and experiment with computational models for automatically identifying such acts from counter-hate arguments.

Model	Variant	Denouncing	Facts	Humor	Hypocrisy	Positive	Question
Roberta	Masked	0.80	<b>0.79</b>	0.70	<b>0.76</b>	<b>0.67</b>	<b>0.86</b>
Roberta	Non-masked	0.76	0.77	<b>0.81</b>	0.69	0.62	0.80
Deberta	Masked	0.76	<b>0.79</b>	0.73	0.71	<b>0.67</b>	0.83
Deberta	Non-masked	<b>0.81</b>	0.78	0.80	0.72	0.66	0.85

Table 3: Macro-F1 scores on validation set for detecting counter-argument type from text.

Model	Variant	Denouncing	Facts	Humor	Hypocrisy	Question	Average
Roberta	Masked	0.45	0.79	0.53	0.38	<b>0.80</b>	0.59
Roberta	Non-masked	0.61	0.78	<b>0.68</b>	<b>0.43</b>	0.74	<b>0.65</b>
Deberta	Masked	0.46	0.78	0.46	0.33	0.69	0.54
Deberta	Non-masked	0.47	0.74	0.55	0.33	0.71	0.56
(Chung et al., 2021a)	Monolingual	0.65	0.84	0.45	0.35	0.73	0.60
(Chung et al., 2021a)	Translated	<b>0.70</b>	<b>0.85</b>	0.56	0.40	0.75	<b>0.65</b>

Table 4: Macro-F1 scores on external test set for detecting counter-argument type from text.

### 3.4.1 Model Architecture

We experiment with pre-trained Roberta (Liu et al., 2019) and Deberta (He et al., 2021) large encoder-based transformer architectures (Vaswani et al., 2017) and expand the encoder’s embedding layer with 6 tokens representing each counter-argument type. The encoder input a hate and counter-hate comment-response pair enclosed by beginning (BOS) and end-of-string (EOS) tokens and separated using two separator tokens. The encoder input is prefixed with the 6 tokens to facilitate the learning representation of each counter-argument type. We pass the encoder representation for the argument type and the BOS tokens through two multi-headed attention layers with four attention heads. We independently apply a linear layer to the resultant encoding of each argument type and predict its logit. The model is trained end-to-end by minimizing the BCE loss function using AdamW optimizer and a learning rate of  $1e-5$ .

Since the Conan dataset only pertains to Islamophobia, we employ a strategy to mask Islamophobic keywords to make the model generalizable to other topics. We computationally curate a set of keywords for each discussion topic in the target dataset. We parse each sentence using Spacy<sup>3</sup> and preserve adjectives, adverbs, interjections, nouns, and verbs that are only present in a single category at most five times as category-specific frequent keywords of a topic. We further expand the keywords using Wordnet by adding related forms and pertainyms and also include their plural forms using inflect<sup>4</sup>. We mask the keywords using a #MASK#

<sup>3</sup><https://spacy.io>

<sup>4</sup><https://github.com/jaraco/inflect>

token and train a masked model variant.

### 3.4.2 Experiments and Results

We use the English split of the Conan dataset for our experiments and include the French/Italian to English translated and paraphrased versions to yield 6,645 examples. We create a train/test set of 6,445/200 instances while ensuring that the counter-arguments present in the test set are unseen in the training set. We experiment with four model variants: masked and non-masked versions of Roberta and Deberta-large and report results in Table 3. We highlight the best scores for each class in bold and observe that (i) The Roberta-based model usually outperforms Deberta with masking and vice-versa without masking. (ii) Except for the classes Denouncing and Humor, masking the inputs yields better results than non-masking. During inference, we create a majority-based ensemble of all four models and only annotate the counter-hate argument in each dialogue turn. We observe a low distribution of examples belonging to Humor and restrict ourselves to the top 5 classes.

To further validate the efficacy of our implementation, we re-train all four model variants on the train/test split by Chung et al. (2021a) and compare our results in Table 4. We highlight the best scores for each class in bold and observe that (i) The Roberta-based model usually outperforms Deberta with and without masking. (ii) Except for the classes Facts and Questions, not masking the inputs yields better results than masking. (iii) Although we achieve the same overall score, our implementation achieves a better score for detecting Humor, Hypocrisy, and Questions. We attribute the differences in our model performance in Tables 3 and 4

to the differences in data curation and size. Chung et al. (2021a) filters out a significant amount of samples and only considers examples that belong to one of the top 5 classes. On the contrary, instead of filtering the data points, we task them as a multi-class binary prediction problem over the 6 most frequent classes.

## 4 Response Generation

We experiment with controllable computational models for generating responses in hateful dialogues. Our objectives are: (i) Assessing the feasibility and utility of using the value and structure-based features for generating and controlling the counter-arguments to hateful speech. (ii) Determining the best feature combination for optimal control over response generation. (iii) Understanding the importance of each of the four sets of features.

### 4.1 Dataset

We use the DialoConan dataset (Bonaldi et al., 2022), which comprises over 3,000 fictitious multi-turn dialogues between a hater and an NGO operator, covering 6 targets of hate: LGBT+, migrants, Muslims, Jews, people of color, and women, and additionally include the hate and counter-hate pairs from the Conan dataset (Chung et al., 2019). We use the computational models discussed in Section 3 to automatically annotate the dialogue corpora, yielding a silver standard dataset.<sup>5</sup>

### 4.2 Model and Training

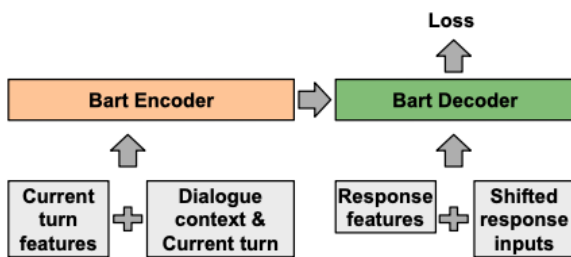


Figure 1: Response Generator Architecture.

Illustrated in Figure 1, we use the transformer encoder-decoder-based Bart-base (Lewis et al., 2020) as the response generator and expand its embedding layer to add the features as tokens. We concatenate the dialogue context and the current turn’s hate speech (query) as the encoder input and additionally prefix the encoder input with the features only present in the current query text and

not in the response. The decoder inputs comprise features that are only present in the current turn response and not in the query text, along with shifted response tokens. The model is trained end-to-end by minimizing the Cross-Entropy loss function using AdamW optimizer and a learning rate of 1e-5. During inference, the encoder input representation is kept similar to training, while the decoder inputs the desired response features and auto-regressively generates a response using beam search with a beam width of 5.

## 4.3 Results and Analysis

We perform extensive automatic and human evaluations to ascertain the usefulness of each category of features and report results.

### 4.3.1 Automatic Evaluations

To determine the best features, we train computational models with different combinations of the value (Val) and structure-based (Struct) features and compare automatic metrics such as corpus BLEU (Papineni et al., 2002), Rouge-L (Lin, 2004), and language modeling perplexity (PPL). We train a model variant without any features as the internal baseline and report our results in Table 5. For each depth of combination (1/2/3 features together) and metric, we highlight (in bold) the best-performing set of features and underline the overall best-performing feature combination per metric.

ID	Type	Features	BLEU	RougeL	PPL
1	Baseline	None	7.83	23.18	7.88
2	Val	big5	<b>8.43</b>	23.76	7.79
3		humVal	7.95	23.12	7.89
4	Struct	argSch	8.25	<b>23.77</b>	7.88
5		argType	8.10	23.37	<b>7.65</b>
6	Val	humVal+big5	7.37	22.72	7.85
7	Struct	argSch+argType	8.24	23.53	<b>7.68</b>
8	Val+ Struct	argSch+big5	<b>8.89</b>	<b>24.60</b>	7.83
9		humVal+argSch	8.10	23.92	7.86
10		humVal+argType	7.34	21.95	7.72
11		big5+argType	7.15	22.31	7.71
12	Val+ Struct	humVal+big5 +argSch	<b>7.52</b>	<b>23.13</b>	7.86
13		humVal+big5 +argType	7.17	22.41	7.69
14		big5+argSch +argType	7.12	22.07	7.68
15		humVal+argSch +argType	6.86	21.78	7.77
16		All	6.75	22.23	<b>7.66</b>

Table 5: Automatic Evaluation Results of Generated Responses for All Feature Combinations.

We observe that (i) When considered independently using the value-based Big-5 personality-

<sup>5</sup>The annotated datasets are strictly used only for research.

based features yields the best BLEU scores, followed by the structure-based argument schemes, and vice-versa for the Rouge-L score. (ii) Considering feature pairs, combining the Big-5 personality and argument schemes produces the best paired and global results. (iii) Among the value-based features, using the Big-5 personality traits generally yields better results than human values. (iv) Among the structure-based features, using the argument scheme generally yields better results than the argument type. (v) Structure-based features generally yield better results than value-based features. However, using both types of features in tandem yields the best results (row 8). (vi) As depicted in Figure 2, combining three features decreases both BLEU and Rouge-L scores below baseline, and using all features together yields the worst BLEU score. Interestingly, the perplexity score generally improves (decreases) with more additional features. We reason that increasing the number of features better partitions the conditional probabilities of the tokens, which enhances language modeling, resulting in a lower perplexity score. However, it also increases the model complexity by adding more degrees of freedom and can be circumvented by additional training examples. We intend to validate this hypothesis as a future research step.

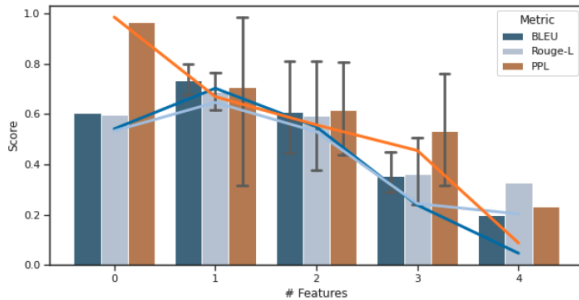


Figure 2: Comparison of Number of features and Normalized (between 0.2 & 1) Score for Different Metrics.

### 4.3.2 Human Evaluations

Employing two evaluators per sample in Amazon Mechanical Turk (AMT), we perform human evaluations on the variants that yield better BLEU scores than the baseline and report our results in Table 6. We define the following three metrics and holistically rate each argument text: (i) **Argumentativeness** (higher better): On a Likert scale of 1 (low) to 5 (high), this indicates the cogency, novelty, and level of engagement of the generated argument. Incorporating proper reasoning,

facts, opinions, and anecdotes, generally makes for a good, convincing, and engaging argument. (ii) **Fluency** (higher better): On a binary scale (0=False, 1=True), this indicates if the generated response is grammatically and syntactically correct and penalizes violations of the rules of the English language other than spelling and punctuation mistakes. (iii) **Hallucination** (lower better): On a binary flag (0=False, 1=True), this indicates if the generated argument is factually and logically sound. It penalizes arguments containing hallucinated facts or illogical propositions which commonly do not make sense in the real world. Figures 6 and 7 (Appendix A) illustrates the evaluation task in AMT.

ID	Type	Features	Arg	Flu	Hal
1	Baseline	None	3.80	0.98	0.06
2	Val	big5	<b>3.85</b>	0.98	0.06
3		humVal	3.67	0.99	0.06
4	Struct	argSch	3.54	0.99	0.05
5		argType	3.78	<b>1.00</b>	<b>0.04</b>
6	Struct	argSch+argType	3.67	0.98	<b>0.03</b>
7	Val+ Struct	argSch+big5	<b>3.80</b>	0.97	<b>0.03</b>
8		humVal+argSch	3.69	<b>0.99</b>	0.04
9		All	3.58	0.98	0.09

Table 6: Human Evaluation Results of Generated Responses for Feature Combinations Better than Baseline.

For each depth of combination and metric in Table 6, we highlight (in bold) the best-performing set of features and underline the overall best-performing feature combination per metric. The annotators agree in 78%, 90%, and 97% cases for argumentativeness, hallucinations, and fluency. We observe that (i) The responses from all model variants are fluent. (ii) For argumentativeness, models incorporating the personality-based features perform better than the baseline, whereas using only argument schemes rates the lowest. Combining personality and argument scheme-based features yields good results, whereas the responses from the model variant using all four features are rated low. (iii) For hallucinations, models incorporating the structure-based features perform better than the baseline and value-based features. Responses from models incorporating a combination of the structure and value-based features have a lower hallucination rate: Combining personality and argument scheme-based features yields good results.

Overall, both automatic and human evaluations indicate that using Big-5 Personality and argument scheme-based features together performs best, demonstrating the need to use value and structure-



based features for generating well-rounded arguments. Further, both types of evaluations indicate an upper bound to the optimal number of feature combinations, where responses using all four types of features are rated low. Table 7 (Appendix A) shares a few examples of the generated counter-hate arguments.

## 5 Conclusion

Motivated to minimize online hate by persuading perpetrators of hate speech contrary to their beliefs, we experiment with ways to generate controllable counter-hate arguments using features derived from psychology, philosophy, linguistics, and social science. We experiment with different feature combinations for generating fluent and persuasive counterarguments. We share computational models for automatically annotating text with such features and a silver-standard annotated hate speech dialog corpora. Our evaluations indicate that a convincing argument should generally appeal to the hateful perpetrator’s personality while incorporating appropriate patterns of reasoning.

## Acknowledgement

We thank the anonymous reviewers for providing valuable feedback on our manuscript. This work is partly supported by NSF grant number IIS2214070. The content in this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding entity.

## Limitations

- We automatically annotate all text with the identified features. The annotations are not validated by human evaluators.
- The hate-speech corpus used in the experiments mainly consists of synthetically curated dialogues. The argument generator’s usefulness is not extensively validated on actual online hate comments.
- The hate-speech dialogue corpus used in the experiments is limited to a set of 6 topics in English. All model performance on any other topics or language is not validated.

## Ethics Statement

We acknowledge that all experiments were performed ethically and purely from an academic point

of view. Although this research revolves around arguments from sensitive topics, the argument generators were not explicitly trained to be discriminatory, exhibit bias, or hurt anyone’s sentiments. Further, any generated text does not reflect the stance of the authors. The human evaluators were appointed and compensated as per the legal norms of Amazon Mechanical Turk. This work is an attempt to improve on counterspeech generation “in the lab”, and is not intended for deployment. It is strictly linguistically motivated and does not compose any psychographic or demographically motivated data profiling, analysis, or segmentation. The proposed methods only utilize diverse sources to derive features based on single turns of short text around specific topics, with the motivation being identifying observable stylistic differences between hateful and counter-hate response text.

## References

- Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. [SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter](#). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
- Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation*, 13(3):429–448.
- Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. [Human-machine collaboration approaches to build a dialogue dataset for hate speech countering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2021a. [Multilingual Counter Narrative Type Classification](#). In *Proceedings of the 8th Workshop on Argument Mining*.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. [CONAN - COUNTER NARRATIVES THROUGH NICHE SOURCING: 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 Tekiroglu, and Marco Guerini. 2020. Italian counter narrative generation to fight online hate speech. In *CLiC-it*.
- Yi-Ling Chung, Serra Sinem Tekiroglu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021c. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.
- Paul T Costa Jr. 1992. Revised neo personality inventory and neo five-factor inventory. *Professional manual*.
- Agustín Manuel de los Riscos and Luis Fernando D’Haro. 2021. Toxicbot: A conversational agent to fight online hate speech. *Conversational dialogue systems for the next decade*, pages 15–30.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
- Damian A Furman, Pablo Torres, Jose A Rodriguez, Lautaro Martinez, Laura Alonso Alemany, Diego Letzen, and Maria Vanina Martinez. 2022. Parsimonious argument annotations for hate speech counter-narratives. *arXiv preprint arXiv:2208.01099*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. *Deberta: Decoding-enhanced bert with disentangled attention*. In *International Conference on Learning Representations*.
- Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*.
- Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. *Identifying the Human Values behind Arguments*. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.
- Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.
- Takahiro Kondo, Koki Washio, Katsuhiko Hayashi, and Yusuke Miyao. 2021. *Bayesian argumentation-scheme networks: A probabilistic model of argument validity facilitated by argumentation schemes*. In *Proceedings of the 8th Workshop on Argument Mining*, pages 112–124, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Ilya Loshchilov and Frank Hutter. 2019. *Decoupled weight decay regularization*. In *International Conference on Learning Representations*.
- François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.
- Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. *The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments*. *CoRR*, abs/2301.13771.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. *Bleu: a method for automatic evaluation of machine translation*. In *Proceedings of the*

- 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Christof Rapp. 2002. Aristotle’s rhetoric.
- Milton Rokeach. 1973. *The nature of human values*. Free press.
- Punyajoy Saha, Kanishk Singh, Adarsh Kumar, Binny Mathew, and Animesh Mukherjee. 2022a. Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. *arXiv preprint arXiv:2205.04304*.
- Sougata Saha, Souvik Das, and Rohini Srihari. 2022b. [Stylistic response generation by controlling personality traits and intent](#). In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 197–211, Dublin, Ireland. Association for Computational Linguistics.
- Sougata Saha and Rohini Srihari. 2023a. [Argu: A controllable factual argument generator](#).
- Sougata Saha and Rohini Srihari. 2023b. [Rudolf christoph eucken at semeval-2023 task 4: An ensemble approach for identifying human values from arguments](#).
- Fillmore H Sanford. 1942. Speech and personality. *Psychological Bulletin*, 39(10):811.
- 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.
- Shalom H. Schwartz. 1994. [Are there universal aspects in the structure and contents of human values?](#) *Journal of Social Issues*, 50(4):19–45.
- Christopher Soto. 2018. *Big Five personality traits*, pages 240–241.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. [Dialogue act modeling for automatic tagging and recognition of conversational speech](#). *Computational Linguistics*, 26(3):339–374.
- 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*.
- Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. [Lamda: Language models for dialog applications](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Jean Wagemans. 2016. Constructing a periodic table of arguments. In *Argumentation, objectivity, and bias: Proceedings of the 11th international conference of the Ontario Society for the Study of Argumentation (OSSA), Windsor, ON: OSSA*, pages 1–12.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation schemes*. Cambridge University Press.
- Steven Windisch, Susann Wiedlitzka, and Ajima Olaghere. 2021. Protocol: Online interventions for reducing hate speech and cyberhate: A systematic review. *Campbell Systematic Reviews*, 17(1).
- Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.
- Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.
- Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. *arXiv preprint arXiv:2106.01625*.

## A Appendix

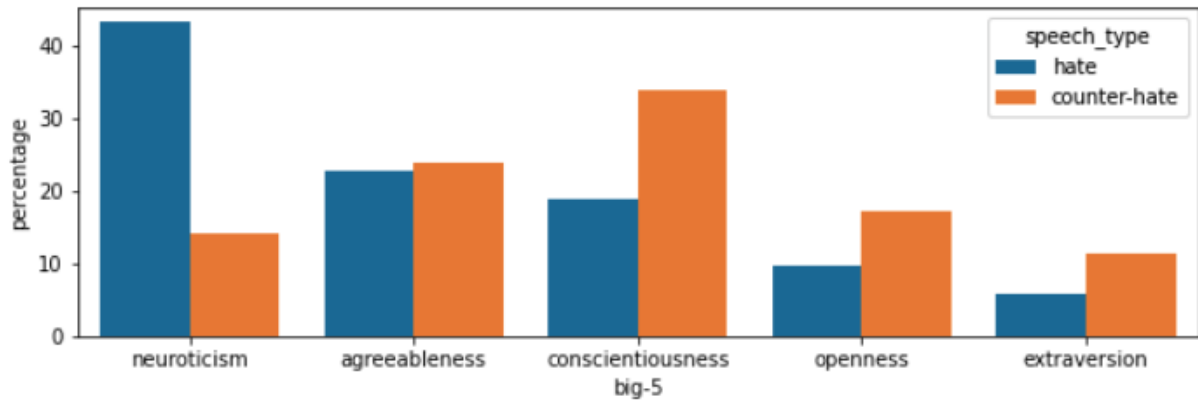


Figure 3: Distribution of Big-5 Personality Traits by type of speech.

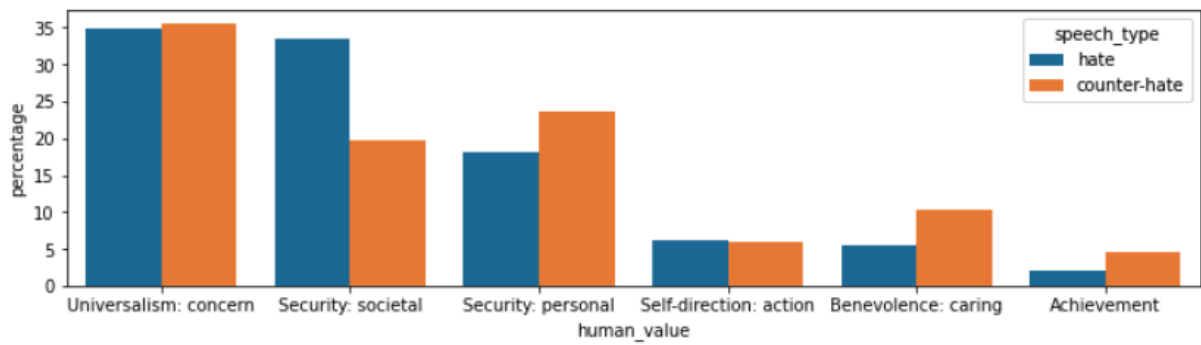


Figure 4: Distribution of Human Values by type of speech.

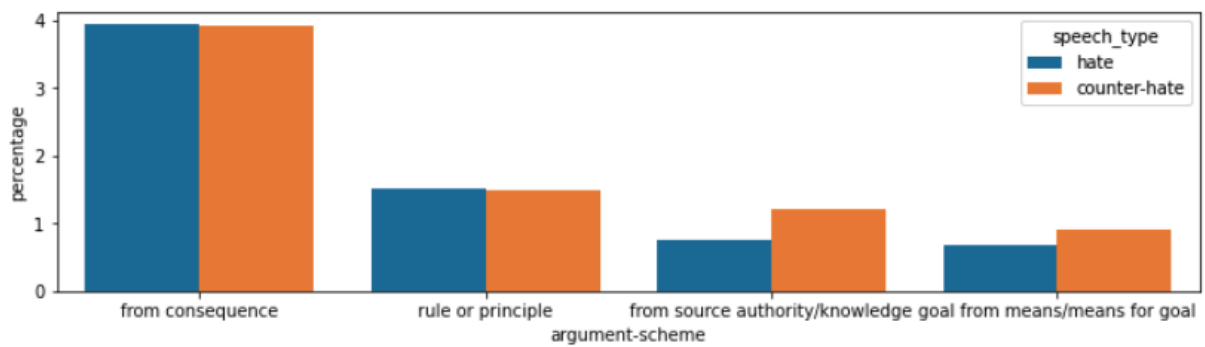


Figure 5: Distribution of Argument Scheme by type of speech. Note: The y-axis is scaled using cubic root.



ID	Context	Baseline	All	argSch+big5
1	<b>&lt;hateSpeech&gt;</b> the hoards of migrants that are coming here are a threat to girls in our country, migrant rape is an epidemic	this is not true. how can you say this about an entire group of people?	i don't see how rape can be considered an epidemic when there are so many migrants coming to our country to work and provide for their families.	rape is always a crime, regardless of who the rapist is. we should focus on stopping it instead of demonising it.
2	<b>&lt;hateSpeech&gt;</b> gays are perverts. the bible teaches that homosexuals are to be despised, a better world is possible only without them. <b>&lt;counterSpeech&gt;</b> what is wrong with being gay for you? i don't think it is wrong to love someone of your same sex. <b>&lt;hateSpeech&gt;</b> gays are different from other people because of their sexual orientation. and deep down in their heart they despise themselves.	why do you say that? do you have any evidence for what you say?	i don't understand why you would say that. homosexuals are not different from other people, they just have different sexual orientations.	gay people don't deserve to be despised for who they are. they are human beings with the same rights as anyone else.
3	<b>&lt;hateSpeech&gt;</b> jews are addicted to their religion and should be isolated. <b>&lt;counterSpeech&gt;</b> even if jews were addicted to their religion, there is nothing wrong in them joining a community of other people, regardless of the fact they're believers. <b>&lt;hateSpeech&gt;</b> they can't join our community because they don't belong to the united states of america.	jews have been part of our country since its creation and have always had a place here. they are our brothers and sisters.	jews have been part of our country since its creation, and have always had a place here. they are our brothers and sisters.	the united states of america is a melting pot of cultures and religions. it is up to us to make sure that jews feel welcome here, regardless of their religion.
4	<b>&lt;hateSpeech&gt;</b> all these migrants are coming to ruin our country, they are all rapists, drug addicts, and don't have any right to be here. <b>&lt;counterSpeech&gt;</b> migration is necessary, to bring in resources and workers that can improve the lives of all. the causes of addiction or sexual violence lie elsewhere. <b>&lt;hateSpeech&gt;</b> they don't improve our lives, immigrants break our laws! they are not welcome here, they are not worth our time.	immigrants come here to work, pay taxes, and have a better quality of life. they are entitled to be here.	they are not breaking our laws, they are just trying to live their life.	immigrants come here to work, pay taxes and have a better life. they are not thieves, they are just trying to live their life.

Table 7: Samples of generated counter-hate arguments from the baseline generator, the best performing generator (argSch+big5), and the generator incorporating all four types of features (all). **OFFENSIVE CONTENT WARNING:** This table contains some examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please be aware that this content could be offensive and cause you distress.

**Task: Evaluate and rate model generated counter-hate speech responses for argumentativeness, in a hateful discussion.**

**Details:** Out of a scale of 1 to 5 (where 1 is lowest and 5 is highest), rate each of the generated responses for argumentativeness. A response is more argumentative if it is convincing and novel. A well rounded and argumentative response generally includes real-world facts, personal anecdotes, well established societal rules, etc. Whereas a weakly argumentative response is mostly bland and generic.

---

**Evaluation Guidelines**

**Ask yourself the following questions while rating each response:**

1. Would you be convinced by the generated argument?
2. Does it include facts, opinions, anecdotes, etc, which makes it a good and engaging argument?
3. Is the argument very generic which makes it less convincing?

**Please Note the Following:**

1. The below arguments contain real world hate speech around minorities, race, religion, gender, etc, which are highly inappropriate. I am sorry that you have to read such hateful and offensive examples.
2. The presentend and generated arguments and text do not reflect or align with the views of the creator of the task.
3. Please select the check-box beside each response to acknowledge that you have thoroughly analyzed the response before rating it.
4. The rating slider starts at 0 by default. However, you are required to give a rating between 1 and 5, where 1 signifies lowest and 5 the highest possible rating. During post-processing at our end, your evaluation WILL BE DISQUALIFIED if we observe ANY of the ratings are still set to the default value of 0.
5. There will be cases where multiple responses contain exactly same (duplicates) or nearly similar (near duplicates) text. In such cases, please be mindful to assign consistent ratings. During post-processing at our end, your evaluation WILL BE DISQUALIFIED if we observe inconsistant scores between duplicate and near duplicate responses.

---

**A few examples**

**Context:** the jews are the reason we have the problems we have today. diversity, immigration, feminism, it's all part of their plan to destroy european people and their culture.

**Response 1:** the jewish people have always been part of european society. diversity, immigration, feminism are not part of a plan to destroy it. they are part of the solution to our problems. The above response is good as it is informative (not generic) and has a positive tone.

**Response 2:** why would jews want to destroy european culture?  
The above response is okayish as it generically challenges the assumptions put forward by the hate speech.

**Response 3:** I do not think that's true  
The above response is acceptable but not the greatest as it only denounces the hate speech.



---

**Evaluate the Following Example:**



**Context:** [hateSpeech] superman is now bisexual, apparently. why is he still called superman?

**Responses**



☐ **Response 1:** bisexuals are not necessarily'superman', they are just attracted to people of the same gender.

○ ————— 0  



☐ **Response 2:** bisexual men are more likely to be bullied and harassed, which is why they are called'superman'.

○ ————— 0  

☐ **Response 3:** bisexual men are more likely to be attracted to men of the same sex than to women of the opposite sex. this is why they are called superman.

○ ————— 0  

☐ **Response 4:** bisexual men are more likely to be bullied and harassed, which is why they're called superman.

○ ————— 0  

**Submit**

Figure 6: Screenshot of Amazon Mechanical Turk Human Intelligence Task for evaluating argumentativeness.

**Task: Evaluate and rate model generated counter-hate speech responses for fluency and factual correctness, in a hateful discussion.**

Details: On a binary scale (Yes/No) rate each argument for: (i) Fluency: Is the generated argument fluent in English? (ii) Falsified Knowledge and Opinion: Does the generated argument factually and logically makes sense?

**Evaluation Guidelines**

**Ask yourself the following questions while rating each response for Fluency:**

1. Is the argument gramatically and syntactically correct?
2. Does it violate the rules of English language? (Note: DO NOT penalize spelling and punctuation mistakes.)

**Ask yourself the following questions while rating each response for Falsified Knowledge and Opinion:**

1. Is the response illogical? Does it commonly make sense in a real world setting?
2. Does it contradict itself?
3. If any factual knowledge is conveyed by an argument, is it correct?
4. If any opinion is expressed by an argument, is it commonly acceptable in the society?

**Please Note the Following:**

1. The below arguments contain real world hate speech around minorities, race, religion, gender, etc, which are highly inappropriate. I am sorry that you have to read such hateful and offensive examples.
2. The presented and generated arguments and text do not reflect or align with the views of the creator of the task.
3. You are getting an unique opportunity to participate in a task that is geared towards social good. The quality of your annotation will directly impact the background research. Hence, please analyze the arguments carefully and mindfully.
4. There will be cases where multiple responses contain exactly same (duplicates) or nearly similar (near duplicates) text. In such cases, please be mindful to assign consistent ratings. During post-processing at our end, your evaluation **WILL BE DISQUALIFIED** if we observe inconsistant scores between duplicate and near duplicate responses.
5. If the presented argument does not express any opinion or shares any facts, please mark **Is Knowledge/Opinion False?** as "No".

**A few examples**

**Context:** the jews are the reason we have the problems we have today. diversity, immigration, feminism, it's all part of their plan to destroy european people and their culture.

**Response 1:** the jewish people have always been part of european society. diversity, immigration, feminism are not part of a plan to destroy it. they are part of the solution to our problems. The above response is fluent and factually correct. Logically the argument makes sense.

**Response 2:** the jewish person is part of european society. diversity, imigration, feminism are not part of a plan to destroy it. The above response is not fluent as it says "jewish person is". Although "immigration" is misspelt as "imigration", we should not penalize it for spelling mistake. Apart from fluency, the presented fact that Jews are a part of european society is true. Hence it factually and logically makes sense.

**Response 3:** the jewish people have always been part of european society. Hence, diversity, immigration, feminism are part of a plan to destroy it. The above response is fluent. The presented fact that Jews are a part of european society is true. However, the statement "Hence, diversity, immigration, feminism are part of a plan to destroy it.", does not make sense and aligns with the hate speech instead of countering it.

**Evaluate the Following Example:**

**Context:** [hateSpeech] superman is now bisexual, apparently. why is he still called superman?

**Responses**

**Response 1:** bisexuals are not necessarily'superman', they are just attracted to people of the same gender.

**Is Fluent?** ☐ Yes ☐ No **| Is Knowledge/Opinion False?** ☐ Yes ☐ No

**Response 2:** bisexual men are more likely to be bullied and harassed, which is why they are called'superman'.

**Is Fluent?** ☐ Yes ☐ No **| Is Knowledge/Opinion False?** ☐ Yes ☐ No

**Response 3:** bisexual men are more likely to be attracted to men of the same sex than to women of the opposite sex. this is why they are called superman.

**Is Fluent?** ☐ Yes ☐ No **| Is Knowledge/Opinion False?** ☐ Yes ☐ No

Submit

Figure 7: Screenshot of Amazon Mechanical Turk Human Intelligence Task for evaluating fluency and hallucination.