IndicCONAN: A Multilingual Dataset for Combating Hate Speech in Indian Context

Nihar Ranja Sahoo, Gyana Prakash Beria, Pushpak Bhattacharyya

CFILT, Indian Institute of Technology Bombay, India
{nihar, gyana, pb}@cse.iitb.ac.in

Abstract

Hate speech (HS) is a growing concern in many parts of the world, including India, where it has led to numerous instances of violence and discrimination. The development of effective counter-narratives (CNs) is a critical step in combating hate speech, but there is a lack of research in this area, especially in non-English languages. In this paper, we introduce a new dataset, IndicCONAN, of counter-narratives against hate speech in Hindi and Indian English. We propose a scalable human-in-the-loop approach for generating counter-narratives by an auto-regressive language model through *machine generation - human correction* cycle, where the model uses augmented data from previous cycles to generate new training samples. These newly generated samples are then reviewed and edited by annotators, leading to further model refinement. The dataset consists of over $\tilde{2}$,500 examples of counter-narratives each in both English and Hindi corresponding to various hate speeches in the Indian context. We also present a framework for generating CNs conditioned on specific CN type with a mean perplexity of 3.85 for English and 3.70 for Hindi, a mean toxicity score of 0.04 for English and 0.06 for Hindi, and a mean diversity of 0.08 for English and 0.14 for Hindi. Our dataset and framework provide valuable resources for researchers and practitioners working to combat hate speech in the Indian context.

Introduction

Hate speech possesses the capacity to inflict numerous harms upon society, including the creation of tension between different groups, contributing to mental health issues, inciting riots, and disrupting peace. Failure to address hate speech can be seen as silently endorsing such behavior, thus fostering a culture of intolerance and worsening its impact on affected communities (Hangartner et al. 2021). However, taking action against hate speech, such as suspending or deleting it, may be perceived as a violation of free speech or as setting a perilous precedent for selective free speech. Instead, counter-narratives emerge as a remarkable solution in this regard. It is viewed as an effective means of combating online hate without compromising freedom of speech (Mathew et al. 2019; Yadav 2018).

Nevertheless, the sheer volume of online hate speech renders effective manual intervention unfeasible, prompting a path of NLP research centered on semi or fully-automated CN generation solutions. In recent times, a range of strategies and datasets for CN compilation have emerged, targeting the data-intensive demands of prevailing generation technologies at the forefront of the field (Chung et al. 2019; Bonaldi et al. 2022).

Our Contributions are:

- A multi-target HS-CN parallel dataset, IndicCONAN, consisting of over 2,500 examples each for Hindi and Indian English covering multiple hate targets in terms of religion, gender, political affiliation, or caste. *To the best of our knowledge, this is the first dataset in the Hindi language for CNs.*
- A framework for generating type-specific CNs using multilingual autoregressive language models, covering various types like *consequences, denouncing, facts, contradiction, counter questions, and positive responses*.
- The effectiveness of CNs was confirmed by language quality, relevance, and diversity metrics. English CNs have a mean perplexity of 3.85, toxicity of 0.04, and diversity of 0.08, while Hindi CNs scored 3.70, 0.06, and 0.14, respectively.

Relevance to Society

Undoubtedly, the impact of hate speech is substantial, with harmful consequences arising when inflammatory language is used to insult, enrage, and even incite violence in extreme cases. Recently, India has experienced a series of violent clashes between religious communities [1] due to the dissemination of divisive and inflammatory speech by certain groups [2]. These incidents not only violate individuals' rights

Warning: *This research paper contains instances and case studies that may be perceived as offensive or targeted. It is important to note that our intention is solely for academic and analytical purposes, and we hold no bias or preference towards or against any specific individual or community.*

[1] https://www.hindustantimes.com/lucknow/muzaffarnagar-riots-fir-against-politicos-for-hate-speeches/story-ymsW4fi9MUtf3Dia3fkz0M.html

[2] https://indianexpress.com/article/opinion/columns/supreme-court-on-mob-lynching-law-against-lynching-case-social-media-whatsapp-rumuors-5265173/

to dignity but also disrupt harmony and tranquility in society. This provides a straightforward rationale for the implementation of laws against hate speech in India.

In Indian law, although the specific term *"hate speech"* is not used, the law does recognize various manifestations of what is commonly understood as hate speech. Within this legal framework, provisions aimed at controlling, curbing, or restraining hate speech are incorporated into our penal code and are upheld under the protection granted by Article 19 (2), which allows for reasonable restrictions on the freedom of speech. It is argued that the diverse array of ways in which Indian laws regulate hate speech creates a challenge in organizing a comprehensive "taxonomy" that adequately encompasses the various descriptions and regulations of hate speech in India (Chinmayi and Nakul 2016). This issue becomes a fundamental consideration each time a study on Indian laws regarding hate speech is initiated. As a result, there is no universally agreed-upon definition of hate speech in India (Online 2017). In India, hate speech is classified as speech that "promotes enmity between groups based on religion, race, place of birth, residence, language, etc." [3], and it can also include "imputations prejudicial to national integration" [4]. Additionally, any references that insult religion or religious beliefs and malicious acts intended to outrage religious sentiments are also considered as hate speech [5]

India's diverse language, caste, race, religion, culture, and beliefs make regulating hate speech a unique challenge. Addressing hate speech based on any of these grounds requires not only punishing the offenders but also restoring the damage caused to the country's secular fabric to prevent further disastrous consequences. Merely relying on legal responses has not been sufficient to deter hate speech incidents. With the rise of online platforms, hate speech has escalated, facilitated by the dissemination of unverified messages, rumors, fake news [6], and deepfakes on social media [7]. These posts aim to instigate and incite violence against specific groups or classes.

Related Work

A counter-narrative is a response to hate speech that utilizes fact-based arguments, counters stereotypes and false information, and alters the viewpoints of people, especially individuals who spread hate speech (Chung et al. 2019). Many studies have found it to be an effective means to not only combat hate but also address the harm that it causes (Yadav 2018; Stroud and Cox 2018; Silverman et al. 2016; Schieb and Preuss 2016). According to a study by Hangartner et al. (2021), counter-narratives enriched with empathy can significantly reduce occurrences of xenophobic hate speech. Mathew et al. (2019) observed that counter-narrative comments on YouTube videos garnered more likes than non-counter-narrative comments, indicating a favorable audience reception towards counter-narratives. According to (Yadav 2018), the most popular counter-narrative pages on Facebook in India are related to 'satirical or religious criticism'.

The increasing popularity of counter-narratives has led to the creation of CN datasets using various methods, including social media scraping, crowd or niche sourcing, and hybrid approaches (Bonaldi et al. 2022; Fanton, Margherita 2021). Chung et al. (2019) developed a multilingual hate speech/counter-narrative dataset on islamophobia.

To expedite the data collection process, Tekiroğlu, Chung, and Guerini (2020); Fanton, Margherita (2021) proposed a hybrid methodology that involved iteratively training a language model to generate pairs of hate speech and counter-narrative, which were then validated by human annotators. Expanding upon this methodology, Bonaldi et al. (2022) introduced a dialog-based data collection approach, that simulates real-life conversations involving multiple exchanges between people. Counter-narratives that effectively refute hate speech through factual information, statistics, and relevant examples are more likely to be accepted. To streamline the process of creating such informative counter-narratives, Chung, Tekiroğlu, and Guerini (2021) developed a generative pipeline that leverages external knowledge acquired through key phrases.

Despite the growing importance of counter-narratives, research on this topic in Indian languages is not available. Capturing Indian cultural and regional contexts requires the use of native languages. In this work, we provide the first multi-class *hate speech-counter narrative parallel dataset* in an Indian language (Hindi) to address hate speech-related issues in India. Moreover, our main contribution is the introduction of a counter-narrative type-based autoregressive model that can identify hate speech and generate a variety of counter-narratives in Hindi.

IndicCONAN Dataset

In this section, we provide an overview of the characteristics of our dataset[8] and elaborate on the methodology employed for annotation and data augmentation. The dataset consists of pairs comprising hate speech and corresponding counter-narratives, where each entry encompasses a hate speech sentence alongside its associated category, complemented by a counter-narrative and its distinctive type. The construction of this dataset was executed via a sequential two-step process. Initially, a manual annotation process was undertaken involving a limited subset of hate speech instances. These instances were meticulously assigned their corresponding categories, and corresponding counter-narratives were annotated. The counter-narratives were further categorized based on their types, such as fact-based, positive tone, contradiction, etc.

Subsequently, we harnessed the human-in-the-loop (HILT) method to iteratively use this annotated set and generate more data points, resulting in larger and more diverse counter-narrative datasets in Hindi and English for the Indian context. These methods will be discussed in detail in

[3]The Indian Penal Code, 1860, Section 153A

[4]The Indian Penal Code, 1860, Section 153B.

[5]The Indian Penal Code, 1860, Section(s) 295A, 298.

[6]https://thewire.in/media/2017s-top-fake-news-stories-circulated-by-the-indian-media

[7]https://factordaily.com/deepfakes-india/

[8]IndicCONAN dataset and scripts used for this work are available at: https://github.com/sahoonihar/IndicCONAN

the following sub-sections.

Hate Speech Datasets

For the initial step, the hate speech instances were collected from two publicly available datasets relevant to the Indian context:-

- **Hostility Detection Dataset**: The dataset (Bhardwaj et al. 2020) comprises 8192 Hindi posts collected from diverse social media platforms (Twitter, Facebook, WhatsApp, etc.). These posts were manually labeled as either hostile or nonhostile. The hostile label includes four dimensions: fake, defamation, hate, and offensive. The hostile category is characterized by four dimensions: fake content, defamation, hate speech, and offensive language. To capture the diverse nature of hostility, a multi-label classification method was utilized, allowing posts to be associated with multiple dimensions simultaneously.
- **The HASOC Fire 2019 Dataset**: The HASOC Dataset (2019), introduced by Mandla et al. (2021), comprises posts from Twitter and Facebook. It is a multilingual dataset containing 7005 posts in English, 5983 posts in Hindi, and 4669 posts in German. The posts were categorized as either HOF (Hate and Offensive) or NOT (Not Hate Offensive). The HOF data was further classified into HATE, OFFN (Offensive), or PRFN (Profanity) categories. For our study, we focused on extracting the Hindi section of the dataset.

Categories of Hate Speech

To enable a more fine-grained classification of hate speech and to explore its relation with various types of counter-narrative, we categorized the hate speech instances into the following types:

- **Caste**: Statements aim to demean, discriminate against, or deride particular castes, or mock the caste system in India.
- **Gender**: Statements humiliating and dehumanizing any specific gender directly or indirectly. Our study focuses exclusively on binary gender categories.
- **Political**: Expressions directed towards political parties, politicians, political actions, policies, and affiliations with the intent to criticize, belittle, or incite negative sentiment.
- **Race**: Statements targeting a person or a group based on their region, state, color, etc.
- **Religion**: Statements that negatively target an individual or a collective based on their religious affiliation or beliefs. This category also includes statements that belittle a religion, its institutions, teachings, and other related aspects.
- **Other**: Statements targeting a person due to other reasons like socio-economic condition, physical appearance, etc.

Types of Counter Narrative

To train a model that can generate diverse and effective counter-narratives for hate speech, a diverse training dataset is necessary. For this, we have decided to consider 6 types of counter-narrative, similar to what has been used in (Chung et al. 2019). Below we discuss these various types:-

- **Positive Response** are statements that involve presenting an optimistic, constructive, or supportive viewpoint in response to a negative or hostile statement. This type of counter-narrative aims to promote understanding, empathy, or positive change by offering an alternative perspective that encourages harmony, cooperation, or a more inclusive outlook.
- **Counter questions** are thought-provoking questions that challenge or probe its underlying assumptions, biases, or implications. This type of counter-narrative aims to stimulate critical thinking and encourage reconsideration of the original statement by prompting reflection on its clarification, validity, fairness, or logical consistency.
- **Denouncing** counter-narratives are statements that involve openly and emphatically condemning a negative or harmful statement, idea, or action. This type of counter-narrative tries to reject and disapprove of the original content by expressing strong opposition by highlighting its negative effects or ethical concerns.
- **Fact-based** counter-narrative provides factual evidence to correct or refute any misperceptions and prevent the spread of misinformation. This approach aims to challenge the accuracy, credibility, or validity of the original content by providing factual data and reliable sources.
- **Warning of consequences** are counter-narratives that inform the user of the potential negative outcomes or repercussions associated with a particular statement, action, or viewpoint.
- **Contradiction and Hypocrisy** are types of counter-narratives that point out any inconsistencies or hypocrisy in the hate-filled statement. By emphasizing the potential consequences, this counter-narrative encourages individuals to reconsider their stance and make more informed choices.
- **Humor and sarcasm** counter-narratives use witty satirical statements to mock or ridicule hate speech. By using sarcasm and comedic techniques, this counter-narrative aims to create a contrasting perspective that exposes the weaknesses of the hate speech while engaging the audience in a thought-provoking and often entertaining manner.

Annotation Strategies for Manual Corpus Creation

By leveraging HILT we followed a two-stage process for dataset creation.

Stage One: From the combined data of the HASOC and Hostility detection datasets, a subset of instances was randomly chosen for annotation of various labels. For each hate speech instance, annotators were tasked with identifying its specific hate type and writing a pair of distinct CNs. The

Post	Hate type	Counter narrative	CN Type
मेरे मन में शंका उठ रही है की तू क्षत्रिय घराने से हो ही नहीं सकता अगर हो सके तो ध्यान में जा कर तेरी माताजी से इतना पूछ लेना की वो कोनसा महान लुंड तेरे पे चढ़ा था की मेरे जैसी अक्कल मंद औलाद को तू ने जन्म दिया.एक थे या अधिक थे या भी पूछ. (I am having doubts whether you belong to a warrior clan, but if you can, ask your mother about the remarkable man who fathered a dim-witted child like you. Inquire if there were one or more contributors to your existence)	Gender	हमारे देश में महिलाओं का बहुत सम्मान किया जाता है और उन्हें देवी के रूप में भी पूजा जाता है। ऐसी संस्कृति में, किसी की माँ के प्रति आपकी अपमानजनक टिप्पणी की सराहना नहीं की जाती है (In our country, women are highly respected and even worshipped as goddesses. In such a culture, your derogatory remarks toward's someone's mother is not appreciated)	Denouncing
चल फुट गद्दार। साला खुद को भारतीय बोलता है और रंडी रोना सिर मुसलमानों के लिए मचाये रहता है। तेरे जात भाई तो जैसे सब दूध के धुले हैं, फरिश्ते हैं? भाग पाकिस्तानी। (Go away, traitor! You call yourself Indian but keep crying for Muslims. Your community members are no angels, are they? Go to Pakistan.)	Religion	ओह मैं समझा। एक खास धर्म की परवाह करना आपको 'गद्दार' बना देता है? मुझे अपनी कट्टरता से अवगत कराने के लिए धन्यवाद। (Oh, I see. Caring for a certain religion makes you a 'gaddar'? Thank you for enlightening me with your bigotry.)	Sarcasm

Figure 1: Examples of paired instances (*hate speech-counter narrtive*) from IndicCONAN corpus. English translations are mentioned in brackets.

counter-narratives were also labeled according to their respective types. The decision to write two distinct CNs per hate speech aimed to introduce diversity into the dataset, allowing the model to learn from multiple perspectives in responding to the same hate speech. However, for a small subset of hate speech, annotators chose to create only a single CN. This decision was not arbitrary; their rationale behind this decision was the concern that introducing other types of CNs might inadvertently escalate tensions caused by that CN instead of effectively countering the hate speech. We also allowed annotators not to write any CN if they were not knowledgeable of the topic that was discussed in the hate speech. For certain hate speech instances, the annotators refrained from producing CNs of three specific types: fact-based, warning of consequences, and contradiction. This was attributed to their limited background knowledge of those topics.

For the above process, we employed two annotators. After receiving the annotations from them, we employed another annotator to validate the correctness of hate speech type labels. Overall, two or more CNs were generated for 91% of hate speech instances, while the remaining 9% received only a single CN. In total, approximately 595 pairs of hate speech CNs were annotated in stage one. Given the complexity of the Hindi language, the initial generation of CNs was carried out in English and subsequently translated into Hindi using NLLB translator[9] (NLLB-Team et al. 2022). We ensured the grammatical and lexical accuracy of the translations through manual verification before incorporating them into the dataset.

In figure 1[10], we provide examples of CNs for each type, showcasing the variety of CNs present in our dataset. For more examples, please refer to the Supplementary material.

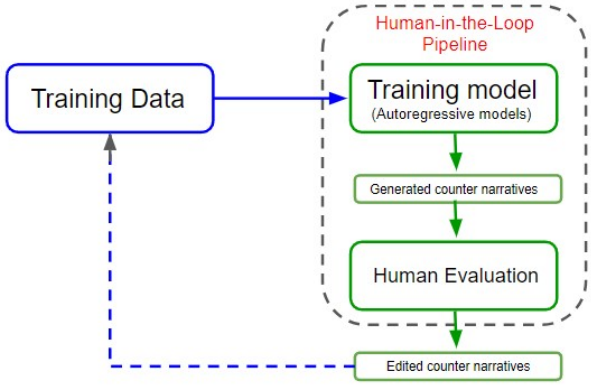

Figure 2: The human-in-the-loop approach for counter-narrative generations.

Stage Two: Following the initial annotation phase, the annotated dataset was divided into three sets: training (475 pairs), validation (40 pairs), and testing (80 pairs). These divisions were established using a stratified approach that took into account both hate speech type and CN type.

Using the approach discussed in we trained both English and Hindi CN generation models using the initial training set. Following this, we utilized the trained English hate speech classifier model to categorize an additional 500 hate speech instances from the combined initial dataset into the six predefined hate speech labels. Subsequently, we subjected these predictions to manual validation for precision. Utilizing the best CN generation model, we conducted in-

[9] facebook/nllb-200-3.3B

[10] As the examples are mentioned in non-roman script, the table is presented as a figure

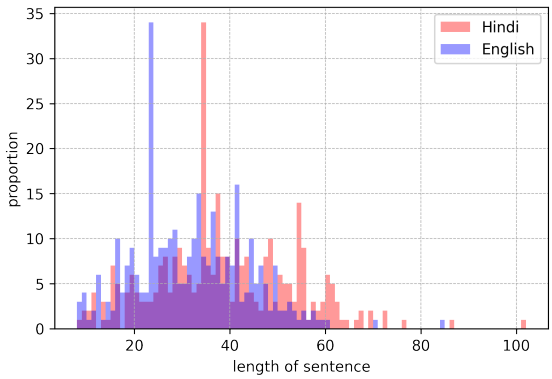

Figure 3: Distribution of word counts for counter-narratives for Hindi and English in training data

ference to create two CNs for each of the 500 hate speech instances, corresponding to two distinct CN types. The selection of these two CN types per hate speech was randomly drawn from the available CN types present in the training set for the respective hate speech category. With the help of annotators, we manually validate those generated CNs and edit them whenever required. The resulting 1000 HS-CN pairs are augmented to the training set. We perform this step separately for English and Hindi languages.

After augmenting these generated pairs, we retrained both English and Hindi CN generation models with the resulting 1475 HS-CN pairs for each language. We then replicated the process outlined in the previous paragraph to generate two CNs for an additional 457 hate speech instances. Once more, we augment the resulting 914 HS-CN pairs to create the final dataset of 2509 counter-narratives for each language.

This corpus creation strategy using HITL is depicted in figure 2.

Data Statistics

In table 1, we have provided the distribution of different categories of hate speech and counter-narratives in our dataset. Additionally, the average word count per instance for each hate speech and counter-narrative category in both English and Hindi is provided in the same table. Considering the potential impact on hate speakers, it is essential to take into account the length of counter-narratives. Shorter sentences may go unnoticed, while longer sentences may become cumbersome to read. Hence, we aimed to maintain the length of counter-narratives between 30 to 50 words. However, sometimes due to the intense nature of certain hate speech instances, the length of counter-narratives might occasionally exceed a 50-word count. Figure 3 illustrates the distribution of word counts for CNs in English and Hindi in our dataset. Notably, the average word count for Hindi is more than English. This can be attributed to the morphological complexity of the Hindi language construct as compared to English.

	Category	Count	#avg_len	
			English	Hindi
HS	Caste	202	44.1	50.2
	Gender	94	33.32	39.88
	Political	250	32.28	38.62
	Religion	413	32.38	38.81
	Race	147	28.43	34.08
	Other	136	29.06	37.41
	Total	1242	33.56	40.03
CN	Positive response	858	33.73	40.92
	Counter question	419	25.75	31.06
	Denouncing	376	25.61	32.90
	Fact-based	222	39.64	46.61
	Warning	265	33.23	40.81
	Contradiction	154	36.6	43.16
	Sarcasm	215	30.18	36.57
	Total	2509	31.52	38.32

Table 1: Distribution of different categories of Hate Speech and Counter-Narratives in IndicCONAN dataset. #avg_len indicates the average number of words per sentence in each category. HS - Hate speech, CN- Counter narrative

Annotation Details

For the initial annotation process, we employed two Indian annotators who are currently pursuing their higher studies. Recognizing the intricacy of the task, we opted to involve three specialized annotators with an understanding of Indian history, culture, and politics rather than resorting to crowdsourcing. We recruited them via a thorough interview process, followed by comprehensive training sessions. During the training, we familiarized them with the nuances of writing counter-narratives and exposed them to publicly available datasets in English. We, initially, gave them 25 hate speech instances to annotate and write CNs. We manually checked their annotations and provided our inputs whenever required. Both the annotators were middle-aged male persons. We employed another middle-aged male to validate their annotations. Three of the annotators helped us verify the model outputs, refine the generations, and edit them whenever required.

We provided detailed guidelines to the annotators with definitions and examples for each haste speech and counter-narrative category. The inter-annotator agreement for the hate speech label of the initial 595 instances was calculated using Cohen's kappa (McHugh 2012). The agreement score was 0.77, which shows very good agreement between annotators.

Experiments and Results

All experiments were run with a single NVIDIA A100 card. All of our implementations use Huggingface's transformer library (Wolf et al. 2020). We use the validation set to decide the best set of hyperparameters. We experiment with learning rates of $1e-5, 2e-5, 3e-5, 4e-5$, epochs of 10, 20, 25, 30, 35, 40, and gradient accumulation steps of 1, 4.

Our dataset consists of four components: $\mathcal{D}_{\mathcal{L}} =$

		Div.	Relevance					LQ.	
	Pipeline	SB1↓	B2↑	B3↑	R1↑	R-L↑	BS↑	PS↓	PER↓
Hindi	**M1**	0.15	0.36	0.20	0.22	0.16	0.69	0.07	3.89
	M2	**0.13**	**0.40**	**0.26**	0.30	0.24	0.72	0.07	3.74
	M3	0.14	0.36	0.19	0.21	0.15	0.69	0.07	4.01
	M4	**0.13**	0.39	**0.26**	**0.32**	**0.25**	**0.75**	**0.06**	**3.70**
English	**M1**	0.10	0.35	0.18	0.17	0.16	0.86	0.05	4.87
	M2	**0.08**	**0.38**	**0.22**	**0.23**	0.21	0.87	0.06	4.20
	M3	**0.08**	0.35	0.19	0.18	0.16	0.86	**0.04**	**3.85**
	M4	**0.08**	0.37	0.21	**0.23**	**0.22**	**0.89**	**0.04**	3.95

Table 2: Performance of different methods using automatic evaluation metrics. The downward arrow ↓ next to the metric name signifies that a lower metric value indicates quality in the generated CN. Conversely, The upward arrow ↑ indicates that a higher metric value corresponds to improved CN quality. The best results are in bold. Div.: Diversity; LQ.: Language Quality; SB1: Self-BLEU-1; B2: BLEU-2; B3: BLEU-3; R1: ROUGE-1; R-L: ROUGE-L; PS: PerspectiveScore; BS: BERTScore; PER: Perplexity. Models represent the different pipelines discussed in .

$\{(h_1, x_1, c_1, y_1), (h_2, x_2, c_2, y_2), ..., (h_n, x_n, c_n, y_n)\}$,
where h_i is a hate speech, x_i is corresponding hate speech category, c_i denotes an annotated/generated CN, and y_i represents the appropriate CN type as annotated by the annotators. The variable L represents the language; it can be either be English or Hindi. When L is Hindi, then both h and c are in Hindi; conversely, the same applies for English. Note that x and y are always in English irrespective of the language of h and c. The objective is to learn a model capable of receiving hate speech h as input and generating a counter-narrative c corresponding to a given h and desired CN type y. Most importantly, our goal is to generate *diverse* and *relevant* CNs for the given HS.

As our main aim is to generate CNs conditioned on the input HS, we used an autoregressive model for the task. Because of the presence of Hindi in our dataset, we decided to use mGPT (Shliazhko et al. 2022) which has both English and Hindi as two of the pre-training languages.

Training Methodology

We experiment with four different training pipelines that are discussed below.

M1: In the first method, we trained the model to learn CNs (c_i) conditioned on only the given HS (h_i) without specifying hate speech type or CN type. Input to the model is a sequence of tokens: $x = \{[BOS], h_i, [SEP], c_i, [EOS]\}$ (without commas). Here $[BOS]$, $[SEP]$, $[EOS]$ are the start token, separator token, and end token respectively. During inference, we provide $[BOS]h_i[SEP]$ sequence as input and stop the model generation once it encounters $[EOS]$ token.

M2: In this method, we trained the model to learn CNs (c_i) conditioned on the given HS (h_i) and CN type without specifying hate speech type. Input to the model is a sequence of tokens: $x = \{[BOS], h_i, [SEP], y_1, [SEP], c_i, [EOS]\}$ (without commas). During inference, we provide $[BOS]h_i[SEP]y_i[SEP]$ sequence as input and stop the model generation once it encounters $[EOS]$ token.

M3: In this method, we trained the model to learn CNs (c_i) conditioned on the given HS (h_i) and hate speech type without specifying CN type. Input to the model is a sequence of tokens: $x = \{[BOS], h_i, [SEP], x_1, [SEP], c_i, [EOS]\}$ (without commas). During inference, we provide $[BOS]h_i[SEP]x_i[SEP]$ sequence as input and stop the model generation once it encounters $[EOS]$ token.

M4: In this method, we trained the model to learn CNs (c_i) conditioned on the given HS (h_i), hate speech type (x_i), and CN type (y_i). Input to the model is a sequence of tokens: $x = \{[BOS], h_i, [SEP], x_1, [SEP], y_i, [SEP], c_i, [EOS]\}$ (without commas). During inference, we provide $[BOS]h_i[SEP]x_i[SEP]y_i[SEP]$ sequence as input and stop the model generation once it encounters $[EOS]$ token.

For all the above methods, during training, we calculate the cross-entropy loss only over the tokens in c_i.

Metrics

For automatic evaluation, we use different metrics to measure the diversity of generated CN and its relevance to the given hate speech. We use the test set to measure the quality of models using these metrics. The test set is similar to $\mathcal{D}_{\mathcal{L}}$ described in the section .

We assess the *diversity* by measuring **Self-BLEU** (Zhu et al. 2018). A lower self-BLEU score corresponds to a higher level of token diversity across all generated CNs. For *relevance*, we evaluate 1) **BLEU** scores (Papineni et al. 2002) between the generated CN and the ground truth CN; 2) **ROUGE** scores (Lin and Hovy 2003) for syntactic between generated and ground truth CN; 3) **BERTScore** (Zhang et al. 2020) for semantic similarity. Relevance metrics indirectly indicate whether the generated CN has a similar topic/context as compared to the given hate speech instance.

We use perplexity (Jansen et al. 2022) as a proxy to measure the linguistic quality of the generated CN. We use the XLMR model[11] to calculate the perplexity of generated

[11] https://huggingface.co/xlm-roberta-base

Post	CN Type	Generated counter narrative
स्मृति ईरानी बिहार में कितने मासूम बच्चो को कन्धा देने गई ? चुनाव ख़त्म! नौटंकी भी ख़त्म! (Smriti Irani went to Bihar to give shoulder to how many innocent children? Election over! The gimmick is over!)	Positive response	राजनेताओं को अपने वादे निभाने के लिए अधिक प्रयास करना चाहिए अन्यथा उन्हें अगला नहीं चुना जाना चाहिए, क्योंकि यह चयन की शक्ति लोकतंत्र के लोगों के पास है। (Politicians should try more to keep their promises otherwise they should not be elected next because this power of selection rests with the people of democracy.)
दिल्ली सरकार में @ArvindKejriwal जैसे नामर्द भरे पड़े हैं। इनके बस का कुछ नही। यह हरामी अब बस ताली पीटने के काम आते हैं। (Delhi government is full of imbeciles like @ArvindKejriwal. Nothing is in their control. These bastards are now only useful for clapping.)	Sarcasm	तो आप दिल्ली सरकार के लिए अच्छा काम करते हैं? तो फिर आपके पास दिल्ली की बेइज्जती क्यों है? हमें यह नहीं ले सकते हैं कि ऐसे आरोपों से और दिल्ली की पुलिस को छुड़ना चाहिए। अन्यथा ऐसी अफवाहें आपको कानूनी परेशानी में डाल सकती (So you do good work for the Delhi government? Then why do you have the insult of Delhi? We cannot accept that Delhi Police should be absolved of such allegations. Otherwise such rumors can land you in legal trouble.)
मुसलमान अल्लाह के सिवाय किसी से नहीं डरते. फिर कश्मीरी मुस्लिम मुह पर कपड़े बांधकर हिजडो की तरह क्यों सामने आते हैं? (Muslims do not fear anyone except Allah. Then why do Kashmiri Muslims appear like eunuchs by tying clothes on their faces?)	Counter question	क्या आप इस बात का कोई प्रमाण दे सकते हैं कि किसी व्यक्ति के कार्य के आधार पर एक पूरे समुदाय को सामान्यीकृत करने के बजाय है? क्या आप अपने दावे का समर्थन करने के लिए सबूत प्रदान कर सकते हैं? (Can you provide any evidence that rather than generalizing an entire community based on the actions of one individual? Can you provide evidence to support your claim?)

Figure 4: Examples of generated counter-narratives for Hindi hate speech instances using our M4 pipeline. Note: English translation of each Hindi output is mentioned in the bracket; not the output of English models.

CNs. Additionally, we use perspective score (Mansourifar et al. 2021) as a measure of the quality of the generated CNs. For each generated CN, the perspective API[12] assigns probabilities for six labels: *toxicity, insult, severe toxicity, identity attack, profanity, threat*. We calculate the average of these six probabilities as perspective scores. A lower perspective score signifies a low level of toxicity within the counter-narrative, aligning with our desired objective.

Results Analysis

In Table 2, we present the numerical values associated with various automated evaluation metrics for both English and Hindi counter-narratives (CNs). The pipeline (M4), conditioned on hate speech (HS), hate speech type, and CN type, exhibits superior performance for both languages. Notably, relevance metrics indicate higher values when the generation is conditioned on CN type (M2 and M4). This outcome is intuitive, as conditioning on CN type imparts the model with ample contextual information for formulating the structure of the counter-narrative.

Diversity scores are quite low across all pipelines also signifies the commendable quality of the initial annotated dataset. The diversity among CNs during the initial annotation stage facilitates the models in acquiring the ability to generate more varied outputs. Additionally, the perspective scores are also significantly low for each pipeline indicating that the generated CNs can be used to counter the hate speech instances effectively. Regarding BERTScore, the relatively low values for Hindi models as compared to English with similar BLEU scores, can be attributed to more word count typically found in Hindi generations.

The table in figure 4 shows the output of M4 pipeline for three Hindi hate speech instances. We provide more model generations in section 3 of our Supplementary material.

Conclusion and Future Works

By introducing IndicCONAN, we encourage the advancement in expanding hate speech-counter narrative research to more low-resource languages and diverse communities. Using multiple automatic evaluation metrics we show the efficacy of our human-in-the-loop pipeline for counter-narrative data creation and generation. We aim to expand our work to further explore the capabilities of other multilingual autoregressive models for CN generations. Moreover, we plan to broaden our reach to include additional Indian languages, reflecting the prevalent preference among Indian users to employ native languages when engaging on social media platforms.

[12]https://perspectiveapi.com/

Ethics Statement

Our research aims to broaden the scope of counter-narrative exploration across diverse cultures and languages. Participants engaged in any facet of this research, whether in data collection, annotation, or evaluation, were diligently furnished with clear and comprehensible information regarding the study's objectives, methodologies, and potential risks. Acknowledging the cultural context of our research, which centers on Hindi and Indian English within the Indian milieu, we have given special consideration to cultural sensitivity. The primary objective of our work is to make a positive contribution to society by addressing the imperative for counter-narrative generation in the Indian context. Furthermore, researchers dedicated to counter-narratives stand to benefit from the dataset we have meticulously collated and the insights gleaned from employing multiple training strategies. We recognize the potential value of these resources in advancing the field of counter-narratives, emphasizing accessibility and ethical responsibility in disseminating our findings for the collective advancement of research in this domain.

Acknowledgements

We would like to thank the anonymous reviewers as well as the AAAI action editors. Their insightful comments helped us improve the current version of the paper.

References

Bhardwaj, M.; Akhtar, M. S.; Ekbal, A.; Das, A.; and Chakraborty, T. 2020. Hostility Detection Dataset in Hindi. arXiv:2011.03588.

Bonaldi, H.; Dellantonio, S.; Tekiroglu, S. S.; and Guerini, M. 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*, 8031–8049. Association for Computational Linguistics.

Chinmayi, A.; and Nakul, N. 2016. Preliminary Findings on Online Hate Speech and the Law in India.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, S. S.; and Guerini, M. 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*, 2819–2829. Florence, Italy: Association for Computational Linguistics.

Chung, Y.-L.; Tekiroğlu, S. S.; and Guerini, M. 2021. Towards Knowledge-Grounded Counter Narrative Generation for Hate Speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 899–914. Online: Association for Computational Linguistics.

Fanton, Margherita. 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.

Hangartner, D.; Gennaro, G.; Alasiri, S.; Bahrich, N.; Bornhoft, A.; Boucher, J.; Demirci, B. B.; Derksen, L.; Hall, A.; Jochum, M.; Munoz, M. M.; Richter, M.; Vogel, F.; Wittwer, S.; Wüthrich, F.; Gilardi, F.; and Donnay, K. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50): e2116310118.

Jansen, T.; Tong, Y.; Zevallos, V.; and Suarez, P. O. 2022. Perplexed by Quality: A Perplexity-based Method for Adult and Harmful Content Detection in Multilingual Heterogeneous Web Data. arXiv:2212.10440.

Lin, C.-Y.; and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Mandla, T.; Modha, S.; Shahi, G. K.; Jaiswal, A. K.; Nandini, D.; Patel, D.; Majumder, P.; and Schäfer, J. 2021. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. arXiv:2108.05927.

Mansourifar, H.; Alsagheer, D.; Shi, W.; Ni, L.; and Huang, Y. 2021. Statistical Analysis of Perspective Scores on Hate Speech Detection. arXiv:2107.02024.

Mathew, B.; Saha, P.; Tharad, H.; Rajgaria, S.; Singhania, P.; Maity, S. K.; Goyal, P.; and Mukherje, A. 2019. Thou shalt not hate: Countering Online Hate Speech. arXiv:1808.04409.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic.

NLLB-Team; Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; Sun, A.; Wang, S.; Wenzek, G.; Youngblood, A.; Akula, B.; Barrault, L.; Gonzalez, G. M.; Hansanti, P.; Hoffman, J.; Jarrett, S.; Sadagopan, K. R.; Rowe, D.; Spruit, S.; Tran, C.; Andrews, P.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Goswami, V.; Guzmán, F.; Koehn, P.; Mourachko, A.; Ropers, C.; Saleem, S.; Schwenk, H.; and Wang, J. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.

Online. 2017. Law Commission of India, Report on Hate Speech, Report no.267. [Online].

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Schieb, C.; and Preuss, M. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ica annual conference, at fukuoka, japan*, 1–23.

Shliazhko, O.; Fenogenova, A.; Tikhonova, M.; Mikhailov, V.; Kozlova, A.; and Shavrina, T. 2022. mGPT: Few-Shot Learners Go Multilingual. arXiv:2204.07580.

Silverman, T.; Stewart, C. J.; Birdwell, J.; and Amanullah, Z. 2016. The impact of counter-narratives. *Institute*

for Strategic Dialogue, London. https://www. strategicdialogue. org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE. pdf–73.

Stroud, S. R.; and Cox, W. 2018. The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, 293–310. Springer.

Tekiroğlu, S. S.; Chung, Y.-L.; and Guerini, M. 2020. Generating Counter Narratives against Online Hate Speech: Data and Strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1177–1190. Online: Association for Computational Linguistics.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771.

Yadav, A. 2018. Counterspeech: An Alternative Policy to Combat Hate Speech in India. *Indian Journal of Law and Human Behaviour*, 4(2): 169–78.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675.

Zhu, Y.; Lu, S.; Zheng, L.; Guo, J.; Zhang, W.; Wang, J.; and Yu, Y. 2018. Texygen: A Benchmarking Platform for Text Generation Models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, 1097–1100. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356572.