Low-Resource Counterspeech Generation for Indic Languages: The Case of Bengali and Hindi

Anonymous ACL submission

Abstract

With the rise of online abuse, the NLP community has begun investigating the use of neural architectures to generate *counterspeech* that can "counter" the vicious tone of such abusive speech and dilute/ameliorate their rippling effect over the social network. However, most of the efforts have so far been primarily focused on English. To bridge the gap for low-resource languages such as Bengali and Hindi, we create a benchmark dataset of 5,062 abusive speech/counterspeech pairs, of which 2,460 pairs are in Bengali, and 2,602 pairs are in Hindi. We implement several baseline models considering various interlingual transfer mechanisms with different configurations to generate suitable counterspeech to set up an effective benchmark[1]. We observe that the monolingual setup yields the best performance. Further, using synthetic transfer, language models can generate counterspeech to some extent; specifically, we notice that transferability is better when languages belong to the same language family.

1 Introduction

The rise of online hostility has become an ominous issue endangering the safety of targeted people and groups and the welfare of society as a whole (Statt, 2017; Vedeler et al., 2019; Johnson et al., 2019). Therefore, to mitigate the widespread use of such hateful content, social media platforms generally rely on content moderation, ranging from deletion of hostile posts, shadow banning, suspension of the user account, etc. (Tekiroglu et al., 2022). However, these strategies could impose restrictions on freedom of expression (Myers West, 2018). Hence one of the alternative approaches to combat the rise of such hateful content is counterspeech (CS). CS is defined as a non-negative direct response to abusive speech (AS) that strives to denounce it by diluting its effect while respecting human rights.

It has already been observed that many NGOs are deploying volunteers to respond to such hateful posts to keep the online space healthy (Chung et al., 2019). Even social media platforms like Facebook have developed guidelines for the general public to counter abusive speech online[2]. However, due to the sheer volume of abusive content, it is an ambitious attempt to manually intervene all hateful posts. Thus, a line of NLP research focuses on semi or fully-automated generation models to assist volunteers involved in writing counterspeech (Tekiroglu et al., 2020; Chung et al., 2020; Fanton et al., 2021; Zhu and Bhat, 2021). These generation models seek to minimize human intervention by providing ideas to the counter speakers that they can further post-edit if required.

However, the majority of these studies are concentrated on the English language. Hence effort is needed to develop datasets and language models (LMs) for low-resource languages. In the past few years, several smearing incidents, such as online anti-religious propaganda, cyber harassment, smearing movements, etc., have been observed in Bangladesh and India (Das et al., 2022a). Bangladesh has more than 150 million people with Bengali as the official language[3], and India has more than 1.3 billion people, with Hindi and English as the official language[4]. So far, several works have been done to detect malicious content in Bengali and Hindi (Mandl et al., 2019; Das et al., 2022b). However, no work has been done to generate automatic counterspeech for these languages.

Our key contributions in this paper are as follows.

- To bridge the research gap, in this paper, we develop a benchmark dataset of 5,062 AS-CS pairs, of which 2,460 pairs are in Bengali and 2,602 pairs are in Hindi. We further label the

[1]We will make our code and dataset public for others

[2]https://counterspeech.fb.com/en/

[3]https://en.wikipedia.org/wiki/Bangladesh

[4]https://en.wikipedia.org/wiki/India

type of CS being used (Benesch et al., 2016b).

- We experiment with several transformer-based baseline models for CS generation considering GPT2, MT5, BLOOM, ChatGPT, etc. and evaluate several interlingual mechanisms.
- We observe that overall the monolingual setting yields the best performance across all the setups. Further, we notice that transfer schemes are more effective when languages belong to the same language family.

2 Related works

This section briefly discusses the relevant work for abusive speech countering on social media platforms and the existing methodologies for CS generation strategies.

Online abuse countering: A series of works have investigated online abusive content, aiming to study the online diffusion of abuse (Mathew et al., 2019a) and creating datasets for abuse detection (Davidson et al., 2017; Mandl et al., 2019; Das et al., 2022b) considering several multilingual languages. In many cases such detection models are used to censor abusive content which may curb the freedom of speech (Myers West, 2018). Therefore as an alternative, NGOs have started employing volunteers to counter online abuse (Chung et al., 2019). Previous studies on countering abusive speech cover several aspects of CS, including defining counterspeech (Benesch et al., 2016a), studying their effectiveness (Wright et al., 2017), and linguistically characterizing online counter speakers' accounts (Mathew et al., 2019b).

CS dataset: So far, several strategies have been followed for the collection of counterspeech datasets. Mathew et al. (Mathew et al., 2019b) crawled comments from Youtube with the replies to that comments and manually annotated the hateful posts along with the counterspeech responses. Chung et al. (Chung et al., 2019) created three multilingual datasets in English, French, and Italian. To construct the dataset, the authors asked native expert annotators to write hate speech, and with the effort of more than 100 operators from three different NGOs, they built the overall dataset. Fanton et al. (Fanton et al., 2021) proposed a novel human-in-the-loop data collection process in which a generative language model is refined iteratively. To our knowledge, no dataset has been built for low-resource languages such as Bengali and Hindi; therefore, in this work, we construct a new benchmark dataset of 5,062 AS-CS pairs for two Indic languages – Bengali and Hindi.

CS generation: Several studies have been conducted for the generation of effective counterspeech. Qian et al. (Qian et al., 2019) employ a mix of automatic and human interventions to generate counternarratives. Tekiroglu et al. (Tekiroglu et al., 2020) presented novel techniques to generate counterspeech using a GPT-2 model with post-facto editing by the experts/annotator groups. Zhu and Bhat (Zhu and Bhat, 2021) suggested an automated pipeline of candidate CS generation and filtering. Chung et al. (Chung et al., 2020) investigated the generation of Italian CS to fight online hate speech. Recently Tekiroglu et al. (Tekiroglu et al., 2022) performed a comparative study of counter-narratives generations considering several transformer-based models such as GPT-2, T5, etc. So far, no work has examined the generation of counterspeech for under-resourced languages such as Bengali and Hindi; therefore, we attempt to fill this critical gap by benchmarking various transformer-based language models.

3 Dataset creation

3.1 Seed sets

Data collection & sampling: To create the CS dataset, we need a seed set of abusive posts for which the counterspeech could be written. For this purpose, we first create a set of abusive lexicons for Bengali and Hindi. We search for tweets using the Twitter API containing phrases from the lexicons, resulting in a sample of 100K tweets for Bengali and 200K for Hindi. The presence of an abusive lexicon in a post does not ensure that the post is abusive; therefore, we randomly sample around 3K data points from both languages and annotate the sample dataset to find out the abusive tweets.

Annotation: We define a post as abusive if it dehumanizes or incites harm toward an individual or a community. It can be done using derogatory or racial slur words within the post targeting a person based on protected attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender (Gupta et al., 2022). Based on the defined guidelines, two PhD students annotated the posts as abusive or non-abusive. Both students have extensive prior experience working with malicious content on social media. After completing the annotation, we remove the conflicting cases and keep the posts labeled as abusive by both annotators. To

measure the annotation quality, we compute the inter-annotator agreement achieving a Cohen's κ of 0.799. Additionally, to increase the diversity of abusive speech in the dataset, we randomly select some annotated abusive speech data points from existing annotated datasets for both Bengali (Das et al., 2022b) and Hindi (Mandl et al., 2019).

3.2 Guidelines for writing counterspeech

Before writing the counterspeech, we develop a set of guidelines that the annotators have to follow to make the writing effective. We define counterspeech as any direct response to abusive or hateful speech which seeks to undermine it without harassing or using an aggressive tone towards the hateful speaker. There could be several techniques to counter abusive speech. Benesch et al. (Benesch et al., 2016a) defines eight strategies that speakers typically use to counter such speech. However, not all of these strategies effectively reduce the propagation of abusive speech. A counterspeech can be deemed successful if it has a positive impact on the hateful speaker. Therefore, the authors further recommended strategies that can facilitate positive influence. As a result, we instructed the annotators to follow the following strategies: *warning of consequences*, *pointing out hypocrisy*, *shaming & labeling*, *affiliation*, *empathy*, and *humor & sarcasm* (see Appendix A for more details).

Annotation process: We use the Amazon Mechanical Turk (AMT) developer sandbox for our annotation task. For the annotation process, we hire 11 annotators, including undergraduate students and researchers in NLP: seven were males, four were females, and all were 24 to 30 years old. Among the 11 annotators, seven are native Hindi speakers, and four are native Bengali speakers. We give them three Indian rupees as compensation for writing each counterspeech. Two expert PhD students led the overall annotation process.

3.3 Dataset Creation Steps

Before starting with the actual annotation, we need a gold-label dataset to train the annotators. Initially, we wrote 20 counterspeech per language, which have been used to train the annotators. We schedule several meetings with the annotators to make them understand the guidelines and the drafted examples.

Pilot annotation: We conduct a pilot annotation on a subset of 10 abusive speech, which helped the annotators understand the counterspeech writing process task. We instruct the annotators to write counterspeech for an abusive speech according to the annotation guidelines. We told them to keep the annotation guidelines open in front of them while writing the counterspeech to have better clarity about the writing strategies. After the pilot annotation, we went through the counterspeech writings and manually checked to verify the annotators' understanding of the task. We observe that although the written counterspeech is appropriate, sometimes, the annotators mislabel the strategy. We consult with them regarding their incorrect strategy labeling so that they could rectify them while doing the subsequent annotations. The pilot annotation is a crucial stage for any dataset creation process as these activities help the annotators better understand the task by correcting their mistakes. In addition, we collect feedback from annotators to enrich the main annotation task.

Main Annotation: After the pilot annotation stage, we proceed with the main annotation task. We gave them 20 abusive speech posts per week for writing the counterspeech. Since consuming a lot of abusive content, can have a negative psychological impact on the annotators, we kept the timeline relaxed and suggested they take at least 5 minute break after writing each counterspeech. Finally, we also had regular meetings with them to ensure that they did not have any adverse effects on their mental health. Our final dataset consists of 5,062 AS-CS pairs, of which 2,460 pairs are in Bengali and 2,602 pairs are in Hindi. We show the distribution of the different types of CS in Figure 1.

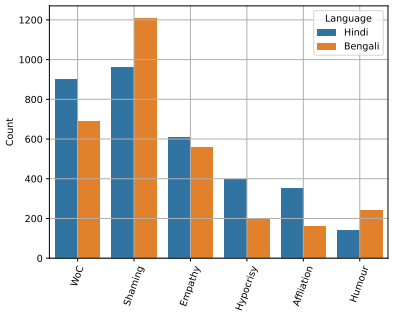

Figure 1: Distribution of the different types of CS based on human annotations.

4 Methodology

4.1 Baseline models

In this section, we discuss the models we implement for the automatic generation of counterspeech. We experiment with a wide range of models.

GPT-2: GPT-2 (Radford et al., 2019) is an unsupervised generative model released by OpenAI only supports the English language. Our focus is to generate counterspeech for non-English language. Therefore to generate counterspeech for Hindi, we use the **GPT2-Hindi** (GPT2-HI) (Parmar) model, and for Bengali, we use the **GPT2-bengali** (GPT2-BN) (Community) model published on Huggingface. (Wolf et al., 2019).

T5-based models: mT5 (Xue et al., 2020), a multilingual variant of T5, is an encoder-decoder model pre-trained on 101 languages released by Google. The mT5 model has five variants, and we use the mT5-base variant for our experiments. For the Hindi language, we also use a fine-tuned mT5-base model, `docT5query-Hindi` (Nogueira et al., 2019), which is trained on a (query passage) from the mMARCO dataset. For Bengali, we also experiment with the `BanglaT5` (Bhattacharjee et al., 2022) model, which is pre-trained with a clean corpus of 27.5 GB Bengali data.

BLOOM: BLOOM (Scao et al., 2022) is an autoregressive large language model developed to continue text from a prompt utilizing highly efficient computational resources on vast amounts of text data, can be trained to accomplish text tasks it has not been explicitly instructed for by casting them as text generation tasks.

ChatGPT: ChatGPT (OpenAI, 2023) is a robust large language model developed by OpenAI, capable of performing various natural language processing tasks such as question answering, language translation, text completion, and many more.

4.2 Interlingual transfer mechanisms

We perform three sets of experiments to check how different models perform under various settings. Specially, we investigate the benefits of using silver label counterspeech datasets to improve the performance of the language models for better counterspeech generation. Below we illustrate the details of these experiments[5].

Monolingual setting: In this setting, we use the same language's gold data points for training, validation, and testing for the counterspeech generation. This scenario generally emerges in the real world, where monolingual datasets are developed and utilized to create classification models, generation models, or models for any other downstream task. Simulating this scenario is more expensive as the gold label dataset has to be built from scratch. In our case, it is the AS-CS dataset.

Joint training: In this setup, while training a model, we combine the datasets of both the Bengali and Hindi languages. The idea is, even though the characters and words used to represent different languages vary, how will these language generation models perform if one wants to create a generalizable model to handle counterspeech generation for multiple languages?

Synthetic transfer: Due to the less availability of datasets in low-resource languages, in this strategy, we experiment with whether resource-rich languages can be helpful if we translate them into low-resource languages and build the generation model from scratch. Further, we experiment that even if some low-resource language datasets are available belonging to the same language community, will it be helpful to generate suitable counterspeeches for other languages? To accomplish this, we use one of the experts annotated English CS datasets (Fanton et al., 2021) (typically constructed with a human-in-the-loop) and translate it into Hindi and Bengali to develop synthetic (silver) counterspeech datasets. Also, we translate the Bengali AS-CS pairs to Hindi and vice-versa to check language transferability between the same language community. In summary, we create the following four synthetic datasets: **EN** $\rightarrow$ **BN**, **HI** $\rightarrow$ **BN**, **EN** $\rightarrow$ **HI**, and **BN** $\rightarrow$ **HI**[6]. We use Google Translate API[7] to perform the translation. Next using the synthetic counterspeech dataset, we build our generation model. In the zero-shot setting (**STx0**), we do not use any gold target instances. In a related few-shot setting, we allow n = 100 and 200 pairs from the available gold AS-CS pairs to fine-tune the generation models. These are called **STx1** and **STx2**.

4.3 Experimental setup

This section describes the training and evaluation approach followed for the language generation models.

4.3.1 Training

All models except ChatGPT were evaluated using the same 70:10:20 train, validation, and test split, ensuring no repetition of AS across sets. For the

[5]For ChatGPT, we only generate CSs in a zero-shot setting. We refrained from fine-tuning due to budget constraints and high computational resource requirements, making it impractical to conduct such experiments.

[6]Languages are represented by ISO 639-1 codes.

[7]`https://cloud.google.com/translate`

Model	Overlap				BERT SC	Diversity	Novelty	Abuse	Human evaluation			
	B-2	B-3	M	ROU		-	-	-	SUI	SPE	GRM	CHO
Bengali												
GPT2-BN	0.053	0.039	0.098	0.166	0.665	0.598	0.807	0.856	3.07	2.75	3.47	0.74
mT5-base	0.117	0.099	0.093	0.178	0.731	0.314	0.637	0.964	3.65	3.07	4.03	**0.90**
BanglaT5	**0.130**	**0.102**	**0.119**	**0.209**	0.724	0.549	0.714	0.972	**3.74**	**3.15**	3.77	0.88
BLOOM	0.093	0.084	0.067	0.139	**0.732**	0.014	0.567	**0.991**	3.73	3.05	**4.42**	**0.90**
ChatGPT	0.024	0.019	0.069	0.094	0.661	**0.850**	**0.914**	0.746	2.58	2.44	3.83	0.615
Hindi												
GPT2-HI	0.101	0.067	0.140	0.244	0.651	0.510	0.778	0.641	2.96	3.12	3.10	0.72
mT5-base	**0.175**	**0.123**	0.133	0.245	**0.715**	0.365	0.674	0.902	3.47	3.15	4.26	0.92
docT5query	0.140	0.103	0.110	0.221	0.698	0.399	0.774	0.608	2.75	2.43	4.16	0.60
BLOOM	0.145	0.108	0.103	0.202	0.712	0.064	0.637	**0.917**	**3.58**	**3.16**	**4.69**	**0.94**
ChatGPT	0.070	0.040	**0.166**	**0.261**	0.673	**0.752**	**0.820**	0.743	2.08	2.48	4.04	0.54

Table 1: Quantitative results of fine-tuned models (monolingual setting) . BERT SC: BERTScore, docT5query: docT5query-Hindi.

Model	Overlap				BERT SC	Diversity	Novelty	Abuse	Human evaluation			
	B-2	B-3	M	ROU		-	-	-	SUI	SPE	GRM	CHO
Bengali												
mT5-base	**0.101**	**0.087**	**0.076**	0.150	0.718	**0.401**	**0.692**	0.967	3.14	**2.71**	4.25	0.85
BLOOM	0.078	0.071	0.070	**0.167**	**0.727**	0.033	0.597	**0.980**	**3.25**	2.67	**4.82**	**0.91**
Hindi												
mT5-base	**0.174**	**0.125**	**0.129**	**0.238**	0.713	**0.391**	**0.695**	0.893	**3.38**	**3.28**	**4.34**	0.80
BLOOM	0.089	0.076	0.073	0.161	**0.717**	0.007	0.593	**0.945**	2.99	2.73	3.94	**0.95**

Table 2: Quantitative results of the fine-tuned models (joint training). BERT SC: BERTScore.

synthetic transfer learning experiments, we split the synthetic datasets into an 85:15 train-validation split. The test set remains exactly the same 20% held out split as earlier. We use 100 and 200 AS-CS gold pairs to further fine-tune the model for the few-shot transfer learning experiments. We make three different random sets for each target dataset to make our evaluation more effective and report the average performance.

Bengali	***Abusive Speech***	দেশ থেকে লাথি মেরে বের করে দেওয়ার পরেও এই নারীর শরম হয় নাই এই নারীর না আছে জ্ঞান না আছে বুদ্ধি তোমার ভতা তেও তুদেও না (Even after being kicked out of the country, this woman has no shame, this woman has neither knowledge nor intelligence, no one even f*ck you)
	Annotated CS	বিদ্বেষপূর্ণ বার্তা ছড়িয়ে আপনি কিছুই লাভ করবেন না বন্ধু। (You won't gain anything by spreading hateful messages dude.)
	mT5-base	আপনি নিজের মুখের ভাষা ঠিক করে কথা বলুন। আমাদের সমাজকে অশ্লীল ভাষায় গালিগালাজ করছেন। (You please speak properly. You are abusing our society with obscene language.)
	BLOOM	আপনার মুখের ভাষা খুবই জঘন্য। মহিলাদের সম্মান করতে শিখুন। (Your language is disgusting. Learn to respect women.)
Hindi	***Abusive Speech***	तुम सुअर की औलादे कुछ भी करे वो दिखाए तो काफिर (No matter what you son of pigs does, if he shows it then he is a Kaafir?)
	Annotated CS	क्या आपको नहीं लगता कि हमें अपने पोस्ट में अश्लील शब्दों का इस्तेमाल नहीं करना चाहिए? क्या यह समाज के लिए अच्छा है? अगली बार कृपया इस तरह के अपशब्दों का प्रयोग करने से पहले विचार करें। (Don't you think we should not use vulgar words in our posts? Is it good for society? Next time please think before using such abuses.)
	docT5query	आपकी पोस्ट को परिभाषित करें। कृपया अपने शब्दों का उपयोग करते समय अधिक सम्मानजनक तरीके से करें। (You rectify your post. Please be more respectful when using your words.)
	BLOOM	किसी व्यक्ति को बदनाम करना और उसका अपमान करना उन्हें आहत कर सकता है। (Slandering and insulting a person can hurt them.)

Table 3: Examples of AS-CS pairs generated by some of the models (monolingual setting).

We use a simple regex-based preprocessing pipeline to remove special characters, URLs, emojis, etc. We limit the maximum length of AS-CS pairs to 400 to include both long and short texts. For the GPT-based and BLOOM models, we follow an autoregressive text generation approach where we separate AS and CS pairs by 'EOS BOS' token to guide the generation to predict suitable CS. For the T5-based models, we use the 'counterspeech' token as the prompt for input and annotated counterspeech as output (more details in Appendix B).

One limitation of ChatGPT is the current inability to fine-tune the model. Given this limitation, our approach to addressing the specific problem of generating counter-speech for abusive language involves crafting well-designed prompts; we aim to generate counter-speech responses for a given abusive speech. We structure the prompts as follows: "*Please write a counter speech in <language name> for the provided abusive speech in <language name>: abusive speech*". Using this prompt, we generate CSs for the test set that was used in all the other models.

4.3.2 CS generation

Following previous research (Tekiroglu et al., 2022), in our experiments, we use the following parameters as default: beam search with five beams and repetition penalty = 2; top-k with $k = 40$; top-p with $p = .92$; min_length = 20 and max_length = 300. We also use sampling to get more diverse generations. We did not need to use any of these parameters for the ChatGPT model. Instead, we passed only the prompt and the AS for which CS had to be generated. We show examples of some generated CSs in Table 3.

4.4 Evaluation metric

We consider several metrics to evaluate various aspects of counterspeech generation. For all metrics, higher is better and the best performance in each column is marked in **bold**, and the second best is underlined.

Overlap metrics: These metrics evaluate the quality of the generation model by comparing the n-gram similarity of the generated outputs to a set of reference texts. We use the counterspeech produced by the various models as candidates and our human written counterspeech as ground truths. To measure how closely the generated counterspeech resembles the ground truth counterspeech, we specifically employ BLEU (**B-2**, **B-3**), METEOR(**M**), and ROUGE-1 (**ROU**).

Diversity metrics: They are used to measure if the generation model produces diverse and novel counterspeech. We employ Jaccard similarity to compute the amount of novel content present in the generated CS compared to the ground truth.

Abusiveness: Finally, to measure the abusiveness of a text, we use `indic-abusive-allInOne-MuRIL` model (Das et al., 2022a) trained on eight different Indic languages in two classes – abusive and non-abusive. We report the confidence between 0-1 for the non-abusive class.

BERTScore: It is an automatic evaluation metric for text generation. Analogously to common metrics, BERTScore (Zhang et al., 2019a) computes a similarity score for each token in the candidate sentence with each token in the reference sentence. However, instead of exact matches, we compute token similarity using contextual embeddings. BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics. We compute BERTscore initialized with the bert-base-multilingual-cased model (Zhang* et al., 2020).

Human evaluation metrics: Despite being difficult to collect, human assessments furnish a more accurate evaluation and a deeper understanding than automatic metrics. Following the previous studies (Chung et al., 2020; Tekiroglu et al., 2022), we also conduct a human evaluation to compare the generation quality of the models under various settings. We use the following dimensions for the assessment of generated counterspeech. **Suitableness** (SUI) measures how suitable the generated CS is in response to the input AS in terms of semantic relatedness and guidelines. **Specificity** (SPE) measures how specific are the explanations obtained by the generated CS as a response to the input AS. **Grammaticality** (GRM) measures how grammatically accurate the generated CS is. **Choose-or-not**(CHO) assesses if the annotators would choose that CS for post-editing and use in a real-life scenario as in the setup suggested by Chung et al. (Chung et al., 2021).

To perform the human evaluation, we randomly select 50 random AS-CS instances from the generated pairs and assign our trained annotators to check the generated CS quality manually.

5 Results

5.1 Performance in the Monolingual Setting

In Table 1, we report the performance in the monolingual setting. We observe that –

For the Bengali language, `BanglaT5` model performs the best across all the **overlapping metrics** (**B-2**: 0.130, **B-3**: 0.102, **M**: 0.119, **ROU**: 0.209), while the mT5-base model performs the second best in terms of BLEU & **ROU** metrics. When considering **BERTScore**, we find that BLOOM achieves the highest score (0.732), closely followed by the mT5-base achieves the second-Highest score (0.731). We notice that BLOOM exhibits the lowest performance in terms of **diversity** (0.014) and **novelty** (0.567), implying that it tends to produce similar responses. In contrast, we observe that ChatGPT exhibited the highest performance, while GPT2-BN exhibited the second-highest score. This indicates that the large language model ChatGPT can generate more diverse counterspeeches compared to the other models. All the models generate mostly non-abusive counterspeeches, with BLOOM achieving the highest score of 0.991 and `BanglaT5` attaining the second-best score of 0.972. In terms of human judgments, the `BanglaT5` model achieves the highest score in terms of **suitableness** & **specificity**. The mT5-base & BLOOM models demonstrate superior performance in the **choose-or-not** metric. In contrast, ChatGPT showed inferior performance in the **choose-or-not** metric, indicating that its responses were not as good to be chosen as counterspeeches in response to an abusive speech.

For the Hindi language, the mT5-base model exhibits the highest BLEU (**B-2**: 0.175, **B-3**: 0.123) while the BLOOM model achieves the second highest score in BLEU (**B-2**: 0.145, **B-3**: 0.108) score. ChatGPT demonstrates the highest performance in terms of METEOR (0.166) score and ROUGE-1 (0.261) score. Regarding **BERTScore**, the mT5-base achieves the highest score (0.715) followed by BLOOM with the second-highest score (0.712). Similar to the Bengali language, we also observe

Model	Overlap				BERT SC	Diversity	Novelty	Abuse	Human evaluation			
	B-2	B-3	M	ROU		-	-	-	SUI	SPE	GRM	CHO
English -> Bengali												
GPT2-BN	0.029	0.025	**0.044**	0.094	0.623	**0.725**	**0.899**	0.672	1.03	1.03	2.05	0.01
mT5-base	0.064	**0.058**	0.042	0.095	**0.689**	0.468	0.863	0.813	1.16	1.13	2.42	**0.12**
BanglaT5	**0.065**	**0.058**	**0.054**	**0.124**	0.676	0.515	0.870	0.828	1.02	1.02	1.61	0.01
BLOOM	0.046	0.043	0.030	0.078	0.658	0.210	0.865	**0.976**	**1.17**	**1.15**	**2.54**	0.10
Hindi -> Bengali												
GPT2-BN	0.026	0.020	**0.067**	**0.140**	0.616	0.522	**0.852**	0.911	**2.32**	**2.04**	3.03	**0.60**
mT5-base	0.080	**0.072**	0.056	0.120	0.702	0.346	0.815	0.981	2.17	1.92	3.07	0.54
BanglaT5	**0.081**	0.070	0.064	0.136	0.691	**0.601**	0.838	0.974	1.70	1.55	2.44	0.32
BLOOM	0.059	0.056	0.037	0.089	**0.705**	0.027	0.825	**0.988**	2.09	1.79	**3.15**	0.36

Table 4: Quantitative results of fine-tuned models for the zero-shot synthetic transfer for Bengali test set. BERT SC: BERTScore.

Model	Overlap				BERT SC	Diversity	Novelty	Abuse	Human evaluation			
	B-2	B-3	M	ROU		-	-	-	SUI	SPE	GRM	CHO
English -> Hindi												
GPT2-HI	0.073	0.049	0.106	0.217	0.626	**0.585**	**0.813**	0.765	1.11	1.09	2.17	0.06
mT5-base	**0.142**	**0.100**	**0.107**	**0.221**	**0.694**	0.501	0.779	0.700	1.25	1.20	3.02	0.16
docT5Query	0.125	0.093	0.089	0.197	0.689	0.462	0.795	0.589	1.33	**1.29**	**3.09**	**0.23**
BLOOM	0.113	0.082	0.092	0.209	0.679	0.307	0.778	**0.794**	**1.32**	1.26	2.95	0.17
Bengali -> Hindi												
GPT2-HI	0.082	0.055	**0.127**	**0.249**	0.647	**0.302**	0.786	0.827	2.40	2.46	3.20	0.04
mT5-base	**0.169**	**0.121**	0.123	0.228	**0.698**	0.179	0.742	0.564	3.46	3.26	4.18	0.58
docT5Query	0.144	0.107	0.101	0.196	0.693	0.123	0.769	0.530	**3.86**	**3.56**	**4.60**	**0.82**
BLOOM	0.097	0.078	0.067	0.159	0.697	0.084	**0.793**	**0.860**	2.48	2.64	3.54	0.12

Table 5: Quantitative results of fine-tuned models for the zero-shot synthetic transfer for Hindi test set. BERT SC: BERTScore, docT5Query: docT5Query-Hindi.

	B-2		M		ROU	
Model	STx1	STx2	STx1	STx2	STx1	STx2
English -> Bengali						
GPT2-BN	0.088	0.027	0.045	0.057	0.100	0.122
mT5-base	0.107	0.114	0.079	0.084	0.171	0.178
Bangla-T5	0.078	0.084	0.063	0.068	0.138	0.155
BLOOM	0.058	0.084	0.054	0.073	0.153	0.167
Hindi -> Bengali						
GPT2-BN	0.027	0.030	0.064	0.073	0.140	0.139
mT5-base	0.102	0.116	0.076	0.087	0.162	0.177
Bangla-T5	0.096	0.103	0.081	0.088	0.161	0.174
BLOOM	0.069	0.069	0.044	0.045	0.103	0.104

Table 6: Few-shot results of the fine-tuned models for the synthetic transfer of EN → BN & HI → BN. Green denotes performance gain (darker denotes larger gain) with respect to STx0 (see Appendix C for EN → HI & BN → HI).

that BLOOM achieves the lowest performance in terms of **diversity** (0.064) and **novelty** (0.637). In contrast, similar to Bengali, ChatGPT demonstrates the highest performance, while `GPT2-HI` exhibits the second-highest score. When considering non-abusiveness, BLOOM and mT5-base achieve good scores. However, GPT2-HI and `docT5query -Hindi` achieve lower scores, indicating that these models often generate abusive speech. Regarding human judgments, we observe that the BLOOM model achieves the highest score in all metrics, while the mT5-base demonstrates the second-highest performance. Similar to Bengali, ChatGPT exhibits poor performance in terms of the **choose-or-not** metric.

Overall, these large language models can generate CSs for low-resource languages. However, the BLOOM model generates less diverse and repetitive counterspeeches in response to abusive speech.

5.2 Performance of the Joint Training

For this experiment, we focus on the mT5-base and BLOOM models due to their capability to handle both Bengali and Hindi languages together. In Table 2, we show the performance of joint training. We see that mT5-base achieves the highest BLEU and METEOR scores for both Bengali and Hindi languages. Similar to the monolingual setting, the BLOOM model exhibits low **diversity** score, indicating that the BLOOM model generates repetitive responses. In terms of human judgment,both models receive high scores for **grammaticality** (GRE) in both Bengali and Hindi, implying their production of grammatically correct responses. However, the **specificity** (SPE) score is less than three for both the models for Bengali and for the BLOOM model for Hindi, indicating that these models produce more generalized responses.

In conclusion, joint training can be employed if a generalizable model is desired to generate counterspeeches for multiple languages.

5.3 Performance of the Synthetic Transfer

In Table 4 & 5, we show the performance of the **STx0** where we synthetically generate AS-CS pairs from the existing dataset. As expected, the performances are less compared to the monolingual setting for both languages. Table 4 reveals that for the Bengali test set, the models trained with **HI → BN**

Bengali	***Abusive Speech***	এগুলো মনে রাখা হবে, মালাউনদের সাথে চূড়ান্ত বুঝাপড়ায় সব কিছুর হিসেব চুকিয়ে দেওয়া হবে, ইনশাআল্লাহ (These will be remembered, all will be settled in the final understanding with the Malauns, Inshallah)
	Zero-shot CS	আপনি যদি এই জাতীয় শব্দগুলি ব্যবহার না করেন তবে এটি আরও সহিংসতার কারণ হতে পারে। (If you do not use such words, it can lead to more violence.)
	Few-shot CS	আপনার এই মন্তব্যে কঠোর ভাষা ব্যবহার করবেন না। এটি আমাদের দেশের জন্য ক্ষতিকারক। (Do not use harsh language in your comments. It is harmful to our country.)
Hindi	***Abusive Speech***	मैं हिन्दू हूँ मुझे तांडव में कोई आपत्तिजनक सीन नहीं लगा स्वरा_भास्कर तू तो रंडी है और रंडी का धर्म सिर्फ पैसा होता है (I am a Hindu, I did not find any objectionable scene in Tandav, Swara_Bhaskar you are a prostitute and the religion of a prostitute is only money.)
	Zero-shot CS	यदि आप इस तरह की टिप्पणियां करते हैं, तो आपको लगता है कि धर्म के आधार पर लोगों का अपमान करना सही नहीं है। (If you make such comments, then you feel that it is not right to insult people on the basis of religion.)
	Few-shot CS	हम सभी का सम्मान करते हैं। कृपया इस पोस्ट को हटा दें। (We respect everyone. Please delete this post.)

Table 7: Examples of AS-CS pairs generated by the mT5-base model in zero-shot & few-shot setting(**STx2**) for **Hi** → **Bn** & **Bn** → **Hi** synthetic transfer. In zero-shot, no gold-label AS-CS pairs were used for training the model.

translated synthetic dataset achieve better scores compared to the **EN** → **BN** translated synthetic dataset. The human evaluation further shows that the generated counterspeeches are of inferior quality for the models trained with **EN** → **BN** translated synthetic dataset. Similarly, in Table 5, we observe that for the Hindi test set, the models trained with **BN** → **HI** translated synthetic dataset achieve better scores compared to the **EN** → **HI** translated synthetic dataset. Human evaluation also indicates an inferior generation of counterspeeches for the models trained with **EN** → **HI** translated synthetic dataset. Among the models trained with **BN** → **HI** translated dataset, we observe `docT5Query-Hindi` and mT5-base models generate counterspeeches with higher scores for human evaluation metrics; however, GPT2-HI and BLOOM show poor performance.

In summary, synthetic transfer schemes exhibit better between Bengali and Hindi languages. This may be attributed by their membership in the Indo-Aryan language family. Table 6 shows the few-shot performance of the synthetic transfer where we add the actual gold AS-CS pairs to fine-tune the models further. Overall we observe adding gold AS-CS gives steady improvements in terms of different overlapping metrics. Hence we recommend instead of developing datasets from scratch, one can use the existing annotated datasets to establish the initial models by performing the synthetic transfer and then fine-tune it for the target language using a small set of gold instances. Table 7 shows some counterspeeches generated in zero-shot & few-shot settings. For the Bengali CS generation, in zero-shot setting, we observe that the CS supports the AS by saying *"if you do not use such words, it can lead to more violence"*[8] – ideally, it should have been the opposite. The generated CS became pertinent in the few-shot setting as it said, *"do not use harsh language in your comments, it is harmful to our country"* – the CS indeed argues that the presence of the offensive word 'Malaun'[9] is harsh and harmful. This shows that the CS generated after the few-shot training is more relevant/semantically consistent.

6 Conclusion

Counterspeech generation using neural architecture-based language models has started gaining attention for interventions against hostility. This paper presents the first attempt at CS generation for the Bengali and Hindi languages, investigating several generation models. To facilitate this, we create a new benchmark dataset of 5,062 AS-CS pairs, of which 2,460 pairs are in Bengali and 2,602 pairs are in Hindi. We experiments with several interlingual transfer mechanisms. Our findings indicate that the overall monolingual setting exhibits the best performance across all the setups. Joint training can be performed if one omnipresent model is beneficial to generate CSs for multiple languages. We also notice synthetic transferability yields better results when languages belong to the same language family.

In future, we plan to explore methods for improving specificity by using various types of knowledge (e.g., facts, events, and named entities) from external resources. Further, we plan to add controllable parameters to the counterspeech generation setup, enabling moderators to customize the counterspeech toward a specific technique we have discussed.

[8]Translated to English.

[9]An offensive word for Hindus.

Limitations

There are a few limitations of our work. First, we have focused solely on generating counterspeech for Bengali and Hindi. Further experimentation should be conducted to address the problem of counterspeech generation in other low-resource languages. By expanding our research to include a broader range of languages, we can better understand the challenges and opportunities in generating effective counterspeech across diverse linguistic contexts. Second, we did not incorporate external knowledge, resources, or facts to enhance the generation of counterspeech. Utilizing such additional information could improve counterspeech generation performance by providing more context and accuracy. Furthermore, while we aim to introduce controllable parameters to customize counterspeech, there are challenges in determining the optimal settings for these parameters. Striking the right balance between customization and maintaining ethical boundaries requires careful consideration and further research.

Ethics Statement

6.1 User privacy

Although our database comprises actual abusive speeches crawled from Twitter, we do not include any personally identifiable information about any user. We follow standard ethical guidelines (Rivers and Lewis, 2014), not making any attempts to track users across sites or deanonymize them.

6.2 Biases

Any biases noticed in the dataset are unintended, and we have no desire to harm anyone or any group.

6.3 Potential harms of CS generation models

Although we observe that these large language models can generate counterspeeches, it is still very far from being coherent and meaningful across the board (Bender et al., 2021). Hence, we do not endorse the deployment of fully automatic pipelines for countering abusive speech (de los Riscos and D'Haro, 2021). Instead, it can be useful as a helping hand to counter speakers in drafting responses to abusive speech.

6.4 Intended use

We share our data to encourage more research on low-resource counterspeech generation. We only release the dataset for research purposes and neither grant a license for commercial use nor for malicious purposes.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016a. Considerations for successful counterspeech. *Dangerous Speech Project*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016b. Counterspeech on twitter: A field study. dangerous speech project.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, and Rifat Shahriyar. 2022. Banglanlg: Benchmarks and resources for evaluating low-resource natural language generation in bangla. *arXiv preprint arXiv:2205.11081*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

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 Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021. Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Flax Community. Flax-community/gpt2-bengali . hugging face. https://huggingface.co/flax-community/gpt2-bengali. Accessed: 2023-04-05.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022a. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. *arXiv preprint arXiv:2204.12543*.

Mithun Das, Somnath Banerjee, Punyajoy Saha, and Animesh Mukherjee. 2022b. Hate speech and offensive language detection in bengali. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 286–296.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In

Proceedings of the International AAAI Conference on Web and Social Media, volume 11.

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.

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

Vikram Gupta, Sumegh Roychowdhury, Mithun Das, Somnath Banerjee, Punyajoy Saha, Binny Mathew, hastagiri prakash vanchinathan, and Animesh Mukherjee. 2022. Multilingual abusive comment detection at scale for indic languages. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Two Hat. 2020. Online moderators: Ten simple steps to decrease your stress. https://www.twohat.com/blog/online-content-moderators-and-reducing-stress/.

Nicola F Johnson, R Leahy, N Johnson Restrepo, Nicolas Velasquez, Ming Zheng, P Manrique, P Devkota, and Stefan Wuchty. 2019. Hidden resilience and adaptive dynamics of the global online hate ecology. *Nature*, 573(7773):261–265.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Binny Mathew, Ritam Dutt, Pawan Goyal, and Animesh Mukherjee. 2019a. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019b. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, 20(11):4366–4383.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

OpenAI. 2023. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed: 2023-04-05.

SURAJ Parmar. Surajp/gpt2-hindi . hugging face. https://huggingface.co/surajp/gpt2-hindi. Accessed: 2023-04-05.

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

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Caitlin Rivers and Bryan Lewis. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

N Statt. 2017. Youtube is facing a full-scale advertising boycott over hate speech. *The Verge*.

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

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

Janikke Solstad Vedeler, Terje Olsen, and John Eriksen. 2019. Hate speech harms: a social justice discussion of disabled norwegians' experiences. *Disability & Society*, 34(3):368–383.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019b. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

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

A Annotation guidelines

A.1 Motivation

Toxic language is prevalent in online social media platforms, presenting a significant challenge. While methods like user bans or message deletion exist, they can potentially infringe upon the principle of free speech. In this task, our objective is to propose a solution that generates counter-speech in response to abusive language, fostering a more constructive online discourse.

A.2 Task

In order to effectively combat abusive language, your task is to craft a well-constructed counter-speech using the recommended strategies outlined in the annotation guidelines. Please ensure that the generated response is clearly marked as a counter-speech, and don't forget to annotate the specific strategy employed to generate the counter-speech. This approach will help us analyze and evaluate the effectiveness of various strategies in addressing abusive language.

A.3 Recommended strategies

There could be several techniques to counter abusive speech. Benesch et al. (Benesch et al., 2016a) distinguish eight such strategies that counter speakers typically use. However, not all strategies help to reduce the propagation of abusive speech. Therefore the author further recommended strategies that can be beneficial to develop positive influence. We discuss these recommended strategies below.

- ***Warning of consequences*** (WoC): In this strategy, the counter speakers often warn of the possible consequences of posting hateful content on public platforms like Twitter. This can occasionally drive the original speaker of the abusive speech to delete his/her source post.

- ***Pointing out hypocrisy***: In this strategy, the counter speaker points out the hypocrisy or contradiction in the user's (abusive) statements. In order to discredit the accusation, the individual may illustrate and rationalize their previous behavior, or if they are persuadable, resolve to evade the dissonant behavior in the future.

- ***Shaming and labeling***: In this strategy, the counter speaker denounces the post as disgusting, abusive, racist, bigoted, misogynistic, etc. This strategy can help the counter speakers reduce the hateful post's impact.

- ***Affiliation***: Affiliation is "... establishing, maintaining, or restoring a positive affective relationship with another person or group". People are more likely to credit the counter-speech of those with whom they affiliate since they tend to "evaluate ingroup members as more trustworthy, honest, loyal, cooperative, and valuable to the group than outgroup members".

- ***Empathy***: In this strategy, the counter speaker uses an empathetic, kind, peaceful tone in response to hateful messages to undermine the abusive post. Changing the tone of a hateful conversation is an effective way of ending the exchange. Although we have little evidence that this will change behavior in the long term, it may prevent the rise of hate speech used at the present moment.

- ***Humor and sarcasm***: Humor is one of the most effective tools used by counter speakers to combat hostile speech. It can de-escalate conflicts and can be used to garner more attention toward the topic. Humor in online environments also eases execration, supports other online speakers, and facilitates social cohesion.

A.4 Dealing with post-annotation stress

We gave the following piece of advice to our annotators – "We understand that the task at hand is

challenging and may have an emotional impact on you. It is important to prioritize your well-being while undertaking these annotations. We strongly recommend taking regular breaks throughout the process. If you find yourself experiencing any form of stress or difficulty, please reach out to the mentors for support. They are there to assist you and may advise you to pause the annotations for a period of 2-3 days to ensure your well-being.

In addition, there is a helpful resource available for you for managing stress in any challenging situation. Please visit https://yourdost.com/ for support and guidance.

We would also wish to provide you with some pointers on dealing with moderator stress. You can find important insights at Hat (2020). In addition, please reach out to your mentors for additional support.

We sincerely appreciate your participation in this annotation task. Your contribution is crucial in furthering our understanding of such societal issues."

B Implementation details

All the models are coded in Python, using the Pytorch library. All training and evaluation have been performed on a Tesla P100-PCIE (16GB) machine with differing batch sizes (GPT2-HI: 1, GPT-BN: 1, mT5-base: 4, `docT5Query-Hindi`: 4, `BanglaT5`: 8, BLOOM: 4) depending on the model architecture. All the models were run up to 50 epochs with Adafactor optimizer (Shazeer and Stern, 2018) having a learning rate of $2e-5$. We save the models for the best validation perplexity score (Zhang et al., 2019b). We also use EarlyStopping patience when validation perplexity decreases by less than $1e-4$.

English->Hindi						
	B-2		**M**		**ROU**	
Model	STx1	STx2	STx1	STx2	STx1	STx2
GPT2-HI	0.088	0.088	0.132	0.131	0.239	0.231
mT5-base	0.156	0.161	0.115	0.117	0.226	0.227
docT5Query	0.142	0.146	0.106	0.111	0.216	0.219
BLOOM	0.111	0.127	0.087	0.096	0.197	0.210
Bengali->Hindi						
GPT2-HI	0.090	0.089	0.138	0.136	0.247	0.238
mT5-base	0.165	0.168	0.123	0.126	0.229	0.235
docT5Query	0.148	0.154	0.106	0.114	0.203	0.214
BLOOM	0.092	0.095	0.062	0.065	0.147	0.155

Table 8: Few-shot results of the fine-tuned models for the synthetic transfer of EN → HI & BN → HI. Green denotes performance gain (darker denotes larger gain) with respect to STx0.

C Synthetic transfer performance

In Table 6, we show the few-shot performance of the synthetic transfer for the EN->HI and HI → BN settings, where we add the actual gold AS-CS pairs to fine-tune the models further.