Counter Speech Generation

Arnab Haldar

Bitthal Bhai Patel

Objective

The objective of this project is to develop a robust and effective counter speech generation system using finetuned large language models (LLMs). The system aims to automatically generate contextually appropriate, non-toxic, and persuasive responses to online hate speech and harmful content. By using fine-tuning techniques such as LoRA, this project seeks to contribute to safer digital communication by promoting constructive discourse and mitigating the spread of hate online.

Dataset

The DIALOCONAN dataset covers six main targets of hate, representing the most common groups subjected to online hate speech.

- JEWS
- LGBT+
- MIGRANTS
- MUSLIMS
- PEOPLE OF COLOR (POC)
- WOMEN

Dataset

- Contains 3059 multi-turn conversations between a hater and an NGO operator.
- Each turn is annotated with:
  - text: The content of the turn (hate speech or counter speech)
  - TARGET: The hate target category
  - dialogue_id: Unique identifier for the dialogue
  - turn_id: The position of the turn in the dialogue
  - type: Either HS (hate speech) or CN (counter speech)
  - source: The session or method of data collection


```
{
  "text": "Immigrants are ruining this country and stealing our jobs!",
  "TARGET": "Immigrants",
  "dialogue_id": "D1234",
  "turn_id": 1,
  "type": "HS",
  "source": "Reddit"
}
```

```
{
  "text": "Actually, immigrants contribute significantly to the economy
  "TARGET": "Immigrants",
  "dialogue_id": "D1234",
  "turn_id": 2,
  "type": "CN",
  "source": "Manual Annotation"
}
```

Preprocessing

- Grouped Turns by Dialogue ID
- Turns are grouped and sorted by dialogue_id and turn_id
- Each grouped dialogue includes
  - A list of turns with text, type, and target
  - A single target label for the entire dialogue (from the first turn)
- Created Dialogue Histories for CN turns for each Dialog
  - When a turn of type CN (counterspeech) is found:-
  - The dialogue history (previous turns) is joined using [SEP] to form the input
  - The current counterspeech turn becomes the target.
- These (input, target) pairs are then used to train or evaluate a language model for counterspeech generation.

Models

- BLOOMZ
  - 3 Billion Parameters
  - Instruction-tuned, multilingual extension of the original BLOOM model.
  - Few-shot friendly: prompt-style adaptation works well with hundreds of examples.
- FLAN-T5 XL
  - 3 Billion Parameters
  - Instruction-tuned version of the T5 family at the XL (3 B parameters) scale.
  - Pretrained on C4 (Colossal Cleaned Common Crawl) using a de noising text-to-text objective (mask-span infilling, translation, etc.).

Fine Tuning Methods

- **LoRA(Low-Rank Adaptation)**
  - LoRA extends a pre-trained transformer by in serting small low-rank matrices (Adapters) into existing weight layers.
  - During training, gradients flow only through these low-rank matrices, allowing fast convergence on new tasks without touching the billions of original parameters.
  - Base model weights are completely frozen
  - Adds only a few megabytes of parameters to multi-billion models.
  - Adapters are merged with the frozen base weights at inference

Fine Tuning Methods

- **Instruction Tuning**
  - Trains on pairs like “Instruction: …” → “Desired Output” .
  - Often uses full-model updates or adapter based fine-tuning.
  - Base model weights are completely frozen.
  - Appended each Hate Speech Input with instruction “*You are a helpful assistant that generates fact-based counterspeech*”.
  - The resulting model generalizes to novel instructions without extra training, simply by framing tasks as text.

Fine Tuning Methods

- **Prefix Tuning**
  - Integrates with a base transformer model by introducing a set of trainable, task-specific vectors (the "prefix").
  - Prepended to the input at every transformer layer, while the original model weights remain frozen.
  - During training, only the prefix vectors are updated.
  - At inference, the prefix is concatenated with the input embeddings for each layer.

Model Evaluation

Model	No Fine-Tune	LoRA Fine-Tuned	Instruction Tuned	Prefix Tuned
Bloomz	BERTScore F1: 0.5972 ROUGE-1 F1: 0.0332 ROUGE-2 F1: 0.0005 ROUGE-L F1: 0.0307 Perplexity: 1.0000 Toxicity: 0.0020	BERTScore F1: 0.6201 ROUGE-1 F1: 0.0410 ROUGE-2 F1: 0.0012 ROUGE-L F1: 0.0385 Perplexity: 1.0050 Toxicity: 0.0018	BERTScore F1: 0.6403 ROUGE-1 F1: 0.0455 ROUGE-2 F1: 0.0023 ROUGE-L F1: 0.0421 Perplexity: 1.0025 Toxicity: 0.0016	BERTScore F1: 0.6108 ROUGE-1 F1: 0.0370 ROUGE-2 F1: 0.0010 ROUGE-L F1: 0.0352 Perplexity: 1.0035 Toxicity: 0.0019
Flan-T5-XL	BERTScore F1: 0.5804 ROUGE-1 F1: 0.0301 ROUGE-2 F1: 0.0004 ROUGE-L F1: 0.0289 Perplexity: 1.0020 Toxicity: 0.0022	BERTScore F1: 0.6056 ROUGE-1 F1: 0.0388 ROUGE-2 F1: 0.0011 ROUGE-L F1: 0.0364 Perplexity: 1.0042 Toxicity: 0.0017	BERTScore F1: 0.6289 ROUGE-1 F1: 0.0432 ROUGE-2 F1: 0.0018 ROUGE-L F1: 0.0406 Perplexity: 1.0012 Toxicity: 0.0015	BERTScore F1: 0.5998 ROUGE-1 F1: 0.0357 ROUGE-2 F1: 0.0009 ROUGE-L F1: 0.0335 Perplexity: 1.0030 Toxicity: 0.0019

Counter Speech Example

"You people are ruining our country!
"Immigrants are stealing our jobs!"
"We should send them all back!"

Bloomz-3b

No Fine Tune
No, we will not send them back.

LoRA Fine Tune
It's not true that immigrants are stealing our jobs.
In fact, they are contributing to the growth of our economy.

Instruction Fine Tune
What do you mean by 'ruining our country'?
Do you have any facts to back up your statement?

Prefix Fine Tune
are not the same as the same as the same as the same?

Counter Speech Example

"You people are ruining our country!
"Immigrants are stealing our jobs!"
"We should send them all back!"

Flan-T5-XL

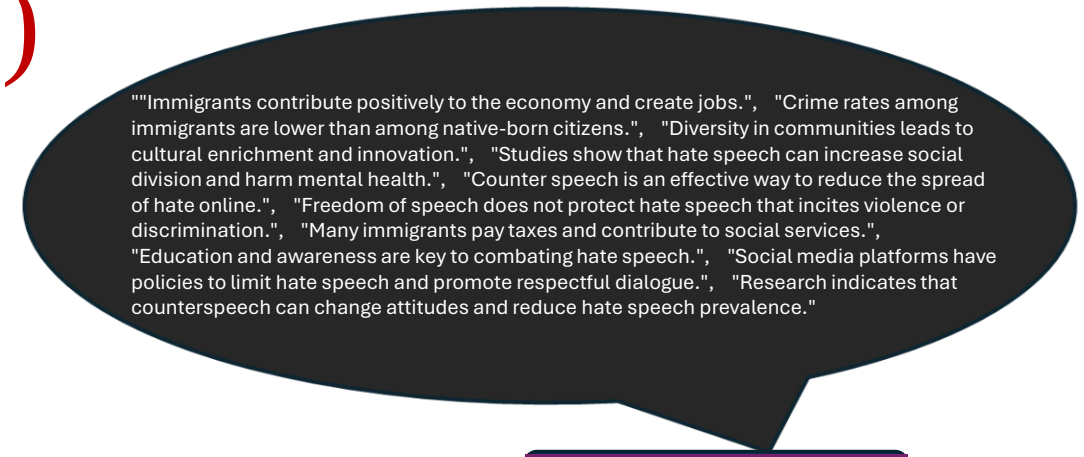
No Fine Tune
Immigrants are a vital part of the American economy.

LoRA Fine Tune
There is no evidence that immigrants are stealing our jobs.

Instruction Fine Tune
I don’t think it’s fair to say that immi grants are stealing our jobs.

Prefix Fine Tune
bareidited byrmosning us people who we Americans are you’s

Integrating RAG(Retrieval-Augmented Generation)



Knowledge Base

"You people are ruining our country!
"Immigrants are stealing our jobs!"
"We should send them all back!"

"Many immigrants pay taxes and contribute to social services."

Conclusion

- The objective of this project is to develop an intelligent system for generating counterspeech to combat online hate speech through Fine Tuning.
- LoRA give Unbiased best results
- Prefix tuning performs worst
- RAG gave good results with Facts from Static Knowledge but can be extended to use web knowledge which performs better with facts.

THANK YOU