ELL884

Project Presentation

Deep Learning for Natural Language Processing

IntentCONANv2 Intent-Specific Counterspeech Generation

Under the supervision of

Prof. Tanmoy Chakraborty

Presented By

Kartik Bansiwal (2021EE30743) & Siddhant Raj (2021MT10932)

Problem Statement

The objective of this project is to build an intent-specific counterspeech generation model using the IntentCONANv2 dataset, which contains approximately 13,000 counterspeech examples. Each counterspeech instance in the dataset is associated with one of four specific intents: informative, denouncing, question, or positive. The task involves generating high-quality counterspeech responses that are not only relevant to the given hate speech input but also aligned with the specified counterspeech intent. The goal is to enable the creation of purpose-driven, context-aware counterspeech that can contribute to more effective and targeted responses in online moderation and discourse.

Dataset

The IntentCONANv2 dataset was provided for this task, containing hate speech–counterspeech pairs annotated with corresponding intent labels. The dataset was split into training, validation, and test sets for experimentation. The final test data released during the competition was completely disjoint from the provided dataset, ensuring an unbiased evaluation of the model's performance on unseen data.

Definitions

1. **Hate Speech:** Language that attacks or demeans a group based on race, religion, ethnicity, gender, sexual orientation, or other identity markers.
2. **Intent:** It refers to the purpose or rhetorical tone of the counterspeech generated in response to a hate speech instance.
3. **Counterspeech:** A direct response that challenges, corrects, or defuses hate speech using positive, informative, or constructive language.
4. **LoRA** (Low-Rank Adaptation): A parameter-efficient fine-tuning method that injects small trainable adapter matrices into a frozen large model, allowing fast, memory-efficient training with strong performance.
5. **T5 (Text-to-Text Transfer Transformer):** A sequence-to-sequence model that frames all NLP tasks as text generation, enabling unified architecture for input → output transformation.
6. **OPT:** OPT-1.3 is a decoder-only causal language model trained to autoregressively generate intent-specific counterspeech from a structured prompt.

Methodology

1. Problem Formulation

- Task: Generate intent-specific counterspeech conditioned on hate speech and intent label
- Intents: Informative, Denouncing, Questioning, Positive
- Input: (intent token + hate speech)
- Output: Counterspeech

2. Data Preparation

- Dataset: IntentCONANv2 (~13k examples)
- Preprocessing:
  - Concatenate intent_token + hatespeech as input
  - Use counterspeech as target
  - Tokenized using T5Tokenizer with truncation/padding
- Train/Validation/Test split: CSVs uploaded manually

Methodology

3. Model Architecture

- Base: T5-Base (Text-to-Text Transfer Transformer)
- Fine-tuned with LoRA adapters (r=8, Q & V projections)
- Generation using beam decoding

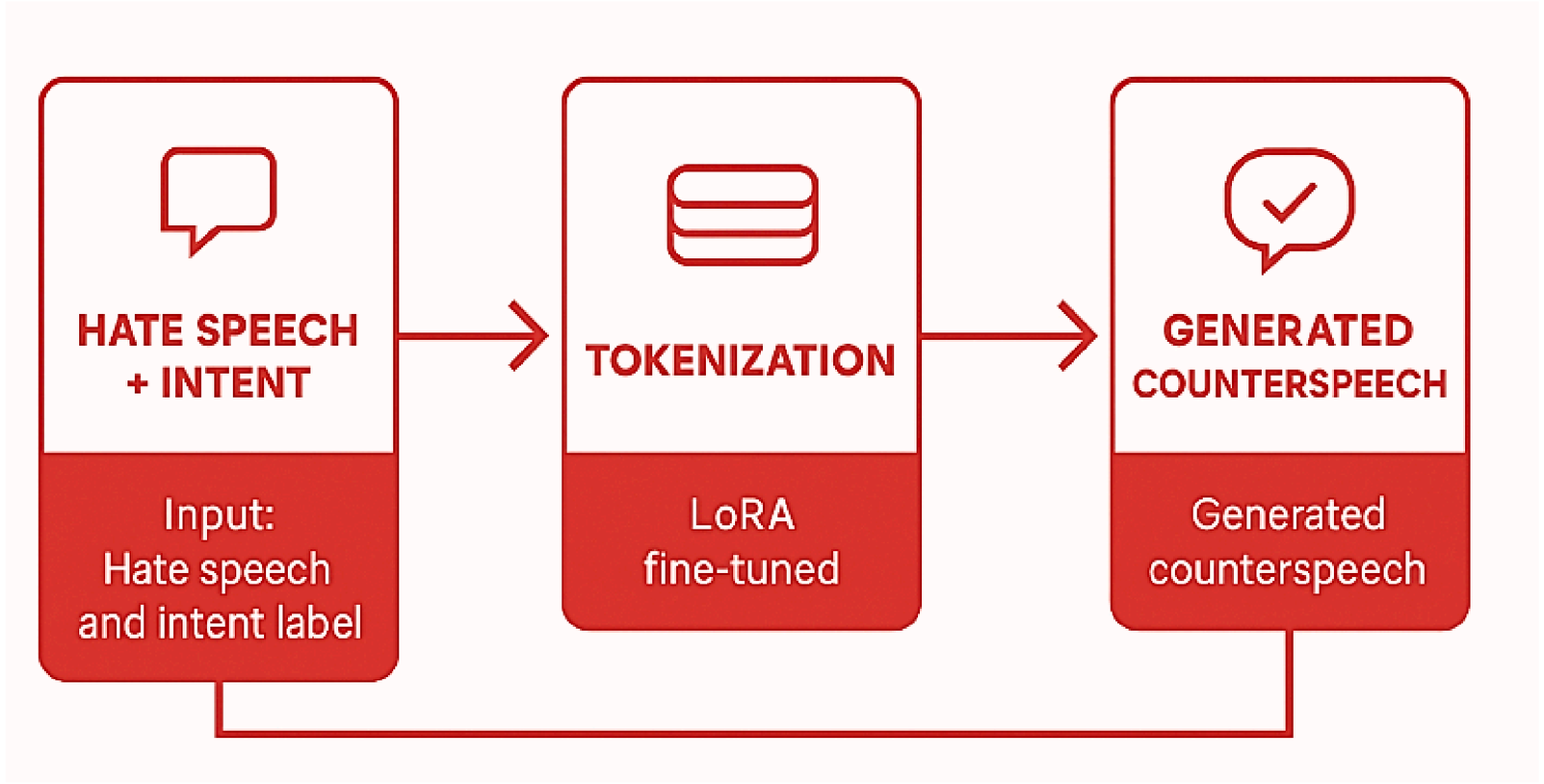
4. Training Strategy

- Seq2Seq finetuning using Transformers Trainer
- Loss: Cross-entropy (with optional label smoothing)
- Evaluation metrics:
    - BLEU (fluency), ROUGE-1/2/L (overlap), BERTScore (semantic similarity)

5. Optimization Enhancements

- Low-Rank Adaptation (LoRA): Efficient finetuning with minimal trainable parameters, Faster training, less overfitting
- Batched inference with decoding optimizations (beam size, length penalty)

Methodology



Training Details - T5

Model: T5-large (770M params) + LoRA adapters

Training Method: Parameter-efficient fine-tuning using LoRA (only adapters trained)

Adapter Settings: r = 8, lora_alpha = 32, applied to query/key projections

Training Data: Combined train.csv + validation.csv from IntentCONANv2 (≈10K samples)

Batch Size: 8 (using gradient accumulation if needed)

Epochs: 3 (initial) → 5 (LoRA phase on merged data)

Learning Rate: 5e-5 (base) → 3e-5 (LoRA tuning)

Tokenizer Max Length: 512 (input), 128 (target)

Training Time: ~24–30 min on Colab (LoRA phase, 3 epochs, 10K examples)

Device: NVIDIA T4 / A100 via Google Colab Pro

Decoding: Beam search (num_beams = 4), max_length = 128, early_stopping = True

Training Details - OPT

Model: OPT-1.3B (1.3B params) + LoRA adapters

Training Method: Parameter-efficient fine-tuning using LoRA (adapters only, base frozen)

Adapter Settings: r = 8, lora_alpha = 16, applied to q_proj and v_proj in decoder blocks

Training Data: Combined train.csv + validation.csv from IntentCONANv2 (≈10K samples)

Batch Size: 2 (with gradient accumulation = 2 for effective batch size = 4)

Epochs: 3

Learning Rate: 2e-4

Tokenizer Max Length: 256 (total input length including prompt + output)

Results

Metric	T5-base	OPT
BLEU	21.6%	23.4%
ROUGE	8.1%	8.7%
BERTScore	0.8702	0.881O

Interpretation:

- OPT consistently outperforms T5-base across all metrics.
- BLEU + ROUGE gains show better lexical and syntactic overlap.
- BERTScore gain confirms improved semantic understanding in OPT generations.

Example

🗣️ Hate: Islam is a disease that is oppressing our beautiful planet.

🎯 Intent: Denouncing

✅ Reference: We denounce this offensive statement. Islam is a diverse religion based on peace, love, and compassion. Blaming an entire faith for the actions of a few goes against understanding and promotes hate.

🤖 Prediction:This statement is baseless and perpetuates harmful stereotypes. Islam, like any other religion, promotes peace, compassion, and respect for all individuals, regardless of their religion.

Conclusion

- Developed two intent-conditioned counterspeech generation models using the IntentCONANv2 dataset.
- Compared T5-base (encoder–decoder) and OPT-1.3B (decoder-only) architectures.
- Applied LoRA adapters to enable efficient, parameter-light fine-tuning on both models.
- Trained on ~10K samples across four intent types: Informative, Questioning, Denouncing, and Positive.
- OPT outperformed T5 across all metrics in fluency and semantic alignment.
- LoRA enabled fast training (~30–45 min) and strong generalization with minimal compute.
- This system demonstrates the potential of intent-aware language models to counter hate speech and promote constructive online dialogue.

ELL884

Project Presentation

Deep Learning for Natural Language Processing

Thank You