Hate Speech Detection Model

Introduction:

Online hate speech is a serious problem that can have a negative impact on individuals and society. Counter-narratives are one way to combat online hate speech. Counter-narratives are stories or arguments that challenge the negative messages of hate speech. They can be used to educate people about the harmful effects of hate speech, to promote tolerance and understanding, and to build a more inclusive society.

Our project generates counter-narratives against online hate speech. The code is written in Python and uses the T5 language model. The T5 language model is a large language model that has been trained on a massive dataset of text and code. The T5 language model can be used to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.

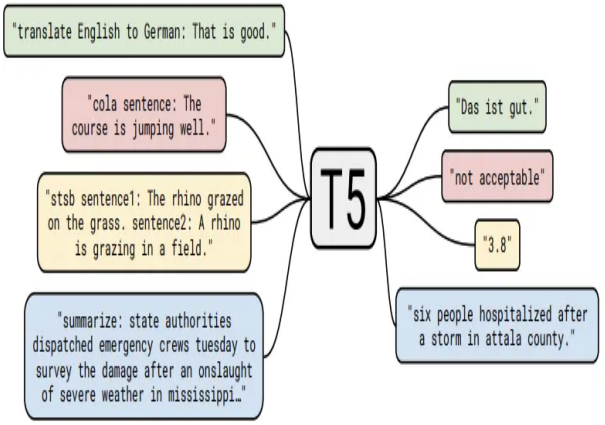
The purpose of this project is to generate counter speech for hate speech using a T5 model. The model is trained using a dataset containing hate speech and its corresponding counter speech.

Dataset:

The dataset used in the provided code is the "Multitarget-CONAN" dataset, which contains instances of hate speech and their corresponding counter-narratives. The dataset was created as part of a research project aimed at developing natural language processing (NLP) tools to automatically generate counter-narratives to hate speech.

The dataset contains the columns Index, Hate speech, Counter narrative, target, and versions. We are only concerned with Hate speech and Counter narrative, so we have chosen only those two columns and rename them into input text and target text respectively. The training and testing split ratio are 0.8 to 0.2 which are of 4002 and 1001 samples.

Method:



T5 (Text-to-Text Transfer Transformer) is a state-of-the-art pre-trained language model that is known for its high performance on various natural language processing tasks, including text classification, question answering, and text generation. The reasons why we have chosen this model are:

1. Large pre-training corpus: T5 was trained on a large-scale corpus of diverse text data, including web pages, books, and articles. This provides the model with a strong understanding of the nuances of natural language, which can help it perform well on a range of downstream tasks.

2. Text-to-text architecture: T5 uses a text-to-text architecture, which means it is trained to map an input sequence to an output sequence. This makes it well-suited for a range of text-based tasks, including text classification, text generation, and machine translation.

3. Fine-tuning capabilities: T5 can be fine-tuned on a wide range of downstream tasks, which makes it a versatile model that can be adapted to a variety of applications. Fine-tuning is a process that involves taking a pre-trained model and further training it on a specific task or dataset, which can help it perform better on that task.

4. State-of-the-art performance: T5 is one of the best-performing pre-trained language models on a range of natural language processing benchmarks, including the GLUE

benchmark and the SuperGLUE benchmark. This suggests that it is a powerful model that can achieve high accuracy on a range of text-based tasks.

Overall, T5 base model is a strong choice for training your dataset due to its large pre-training corpus, text-to-text architecture, fine-tuning capabilities, and state-of-the-art performance on a range of natural language processing tasks.

Training:

[2505/2505 14:33, Epoch 5/5]

Step	Training Loss	Validation Loss
100	2.996900	1.873150
200	1.925200	1.802232
300	1.849500	1.767950
400	1.804900	1.746628
500	1.798900	1.733015
600	1.723500	1.721706
700	1.766600	1.713196
800	1.740700	1.705538
900	1.708800	1.700607
1000	1.690300	1.694980

The T5 Hugging Face model was trained using the training dataset and validated using the validation dataset. The model was configured using Training Arguments to specify the output directory, number of epochs (5 for our model), batch sizes, and logging and evaluation strategies. The Trainer function was used to train the model with the specified parameters.

Validation:

	input_text	target_text	predicted_text
414	migrants steal our jobs	When we give refugees and asylum seekers the o...	The migrants who work here steal our jobs, and...
1815	Persons with learning difficulties should be d...	Even if individuals with learning difficulties...	The education system is designed to help peopl...
642	Many philosophers argue that humans are not bo...	This is a so-called 'argument from authority' ...	Human beings are born equal, and they are born...
1056	LGBTs are a drain on society.	Sexual abuse, physical abuse and even the lack...	I don't think that LGBT people are a 'drain'. ...
4507	Mentally disabled folks are gullible and don't...	A person can have a diagnosable mental disorde...	Mentally disabled people are not gullible, the...
...	...	...	...

Using the tokenizer, we have generated the counter speech to hate speech of the validation dataset and added it as a predicted text column to the validation dataset.

The predicted text and target text are converted into NumPy arrays. Using the Text Blob library, we compute the sentiment polarity and store it in the list. The polarity of the text was calculated as a score between -1 and 1, where negative values represent negative sentiment and positive values represent positive sentiment. The accuracy obtained using T5 base is 0.835.

```
Accuracy Score: 83.51648351648352
```

Conclusion:

As online hate content rises massively, responding to it with counter-narratives as a combating strategy draws the attention of international organizations.

Although a fast and effective responding mechanism can benefit from an automatic generation system, the lack of large datasets of appropriate counter-narratives hinders tackling the problem through supervised approaches such as deep learning.

Team Contribution:

Team Member	**Contribution**
Greeshma Naga Lakshmi Palisetty,Nandanandan Chowdary Tagirisa	Data Collection and preprocessing
Veenus Gollapalli,Akhil Sunkara	Model Training andCounter Speech Generation
Yashwanth Devireddy,Bhavitha Gorrepati	Model Selection and Validation