GENERATING COUNTER-SPEECH AGAINST ONLINE HATE-SPEECH

Introduction:

The popularity of social media has increased global communication and connectivity, but it has also given a platform for hate speech. Therefore, it has become crucial to combat hate speech in order to sustain a vibrant online community. Using counter-narratives with machine learning models like T5 can combat online hate speech. T5 is versatile and produces high-quality text. It can automatically create new counter-narratives when trained on a dataset of hate speech and counter speech. This reduces the time and effort needed to construct counter speech manually. Python is a suitable programming language for creating a system using T5 to generate counter-narratives against online hate speech.

Conan Dataset:

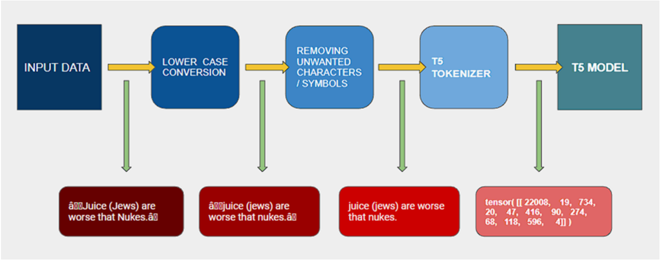
The CONAN dataset was created through crowdsourcing to gather responses to hate speech scenarios. It contains over 5003 high-quality counter-narratives tagged by the type of hate speech addressed. The dataset's diversity and quality makes it a valuable resource for training and evaluating machine learning models.

Method (T5 Model Comparison)

Transformers are models that can translate text, write poems and op-eds, and even generate computer codes. The machine learning models like BERT, GPT-3 and T5 are based on transformers. Transformers are a type of neural network architecture which are a very effective type of model for analyzing complicated data, types like videos, audios, and text. Until transformers were introduced, we used RNNs but they had a lot of problems.

T5 (Text- to-Text Transfer Transformer) methodology is an advanced NLP model created by Google that utilizes a text-to-text transfer learning approach. It has the capability to perform a multitude of NLP tasks and is highly adaptable, as it can be trained on different types of datasets. The major reasons behind choosing this are as follows: (1) Versatility: T5 is flexible and can perform a wide range of NLP tasks (2) Large-scale: T5 has been trained on a massive amount of text data, allowing it to develop a deep understanding of language and context. (3) High-quality responses: T5 is a powerful language model that can generate high-quality responses to various tasks. (4) Rapid prototyping: allows developers to quickly test and refine their ideas. (5) Open-source: allows developers to modify and customize its code to build applications tailored to their specific needs.

Data Processing:


INPUT DATA
LOWER CASE CONVERSION
REMOVING UNWANTED CHARACTERS / SYMBOLS
T5 TOKENIZER
T5 MODEL
â��Juice (Jews) are worse that Nukes.â��
â��juice (jews) are worse that nukes.â��
juice (jews) are worse that nukes.
tensor([[22008, 19, 734, 20, 47, 416, 90, 274, 68, 118, 596, 4]])

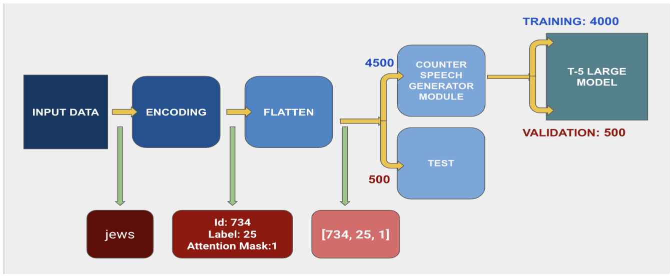
Before feeding the input data to the T5-based model, it undergoes a few preprocessing steps, which are implemented in the pre_processing() function. The pre_processing() function converts all text to lowercase using the .lower() method. The clean() function from the cleantext library with the no_emoji argument set to True removes all emojis from the input text .

Next step is Encoding. The CounterSpeechDataset class is responsible for performing tokenization, padding, and creating PyTorch tensors for the input and target text data. CounterSpeechDataset class is used to preprocess the text data by breaking down the input text into a sequence of tokens, which are then encoded as numerical values. This enables the T5-based model to process the input data and learn the relationships between the tokens in the sequence.

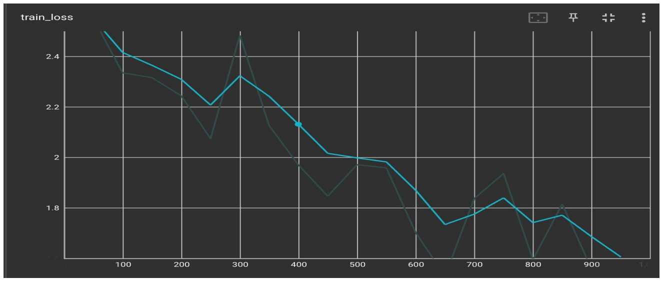
The tokenized text is then padded to ensure that all sequences are the same length. To define the max length of the source and target padded sequences we did data exploration of the hate speech and counter speech token count distribution to get the required values and truncate the Hate Speech input. Finally, the tokenized text is converted into PyTorch tensors and returned as inputs to the machine learning model.

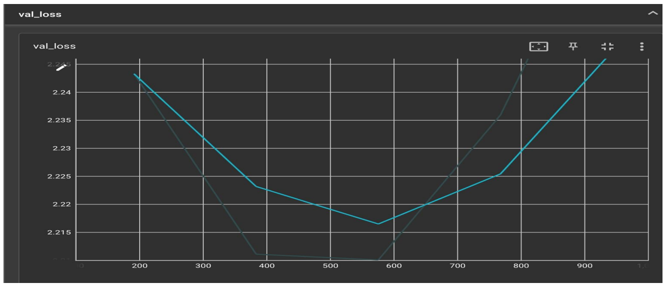
Model Pipeline and Training:

The dictionary returned by the Encoding follows the format required for custom Hugging Face Dataset. It consists of input ids (tokenization of input hate speech), labels (tokenization of input counter speech), Attention-mask (purpose is to mark whether each word is in use or not and is initialized to all one). We then flatten the entire input into a single list. The entire CONAN data set consists of 5000 data points where we split it into 4500 data points for Training (seen data) and and 500 data points for Testing (unseen data). Out of the 4500 we split it to 4000 data points for Training and 500 data points for Validation and saving the best model checkpoint.



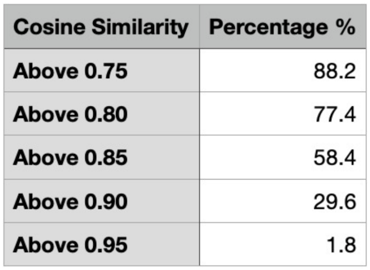
The processed DataSet is passed to the DataLoader as input since pytorch training module requires this as an iterable for model training. From the setup of the CounterSpeechGeneratorModule DataLoader, it returns 2 objects for Train Data Loader and Test Data Loader. We use the model “T-5 Large” as our base pretrained model and apply our training dataset for Transfer Learning. We use the Batch size to be 20 so that our model can converge to global minima faster than by keeping batch size smaller. We keep number of epochs to be 20 as we see that via the validation loss that the best model checkpoint is picked well before 5 epochs as the training loss keeps diminished but the validation loss first decreases and then starts rising up showing that model starts overfitting after 5 epochs. We start with Learning rate of 0.0001 and we also use Adam Optimiser





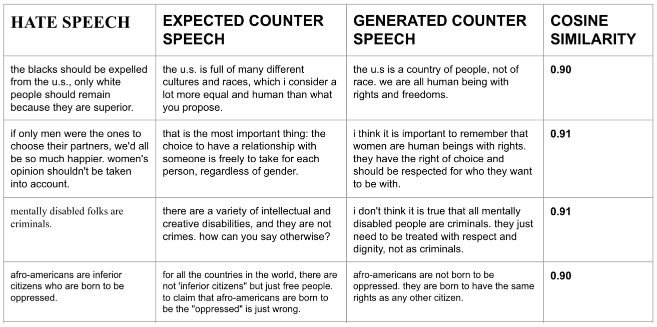
Model Results :

Out of 500 data points on the test set, we used “Cosine similarity” as a metric to measure semantic similarity between given counter speech and generated model output counter speech using the spacy package. We achieve cosine similarity above 0.75 for almost 90% of the test set. Here is the distribution.



Cosine Similarity	Percentage %
Above 0.75	88.2
Above 0.80	77.4
Above 0.85	58.4
Above 0.90	29.6
Above 0.95	1.8

Here are some of the sample model generated answers.



HATE SPEECH	EXPECTED COUNTER SPEECH	GENERATED COUNTER SPEECH	COSINE SIMILARITY
the blacks should be expelled from the u.s., only white people should remain because they are superior.	the u.s. is full of many different cultures and races, which i consider a lot more equal and human than what you propose.	the u.s is a country of people, not of race. we are all human being with rights and freedoms.	**0.90**
if only men were the ones to choose their partners, we'd all be so much happier. women's opinion shouldn't be taken into account.	that is the most important thing: the choice to have a relationship with someone is freely to take for each person, regardless of gender.	i think it is important to remember that women are human beings with rights. they have the right of choice and should be respected for who they want to be with.	**0.91**
mentally disabled folks are criminals.	there are a variety of intellectual and creative disabilities, and they are not crimes. how can you say otherwise?	i don't think it is true that all mentally disabled people are criminals. they just need to be treated with respect and dignity, not as criminals.	**0.91**
afro-americans are inferior citizens who are born to be oppressed.	for all the countries in the world, there are not 'inferior citizens" but just free people. to claim that afro-americans are born to be the "oppressed" is just wrong.	afro-americans are not born to be oppressed. they are born to have the same rights as any other citizen.	**0.90**

Contributions:

Hitesh Atluri	Data collection followed by cleaning and preprocessing.
Sarthak Gupta	Writing entire Model Training pipeline & code , Result Metrics research, ppt and report
Sravya Mareedu	Fine-tuned the model, trained it to produce high-quality counter-speech.
Tharun Reddy	Tokenized dataset and prepared it for fine-tuning, ppt and report.
Monisha Ramu	T5 model comparison report writing and ppt generation, Data Preprocessing.
Satya Pradeep	Identifying suitable dataset for the model,Data Preprocessing.
Akhil Shashi	Exploratory Data Analysis and writing Data Preprocessing and Cosine Similarity code.