**Toxicity Detection and Alternative Phrasing**

**1. Introduction**

**Problem Statement:**

**Goal:** Identify and mitigate toxic, hateful, or harassing comments.

**Objectives:**

- Build a model to classify comments as `toxic` or `neutral`.

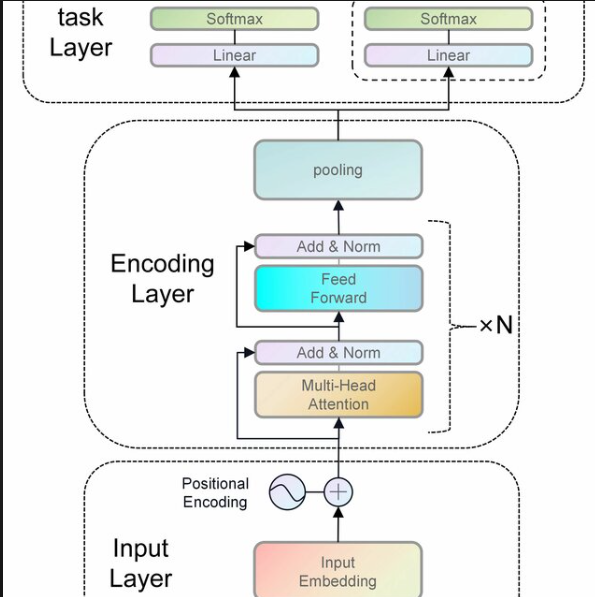
- Generate alternative phrasings for toxic comments.

**2. Model Architecture and Methodology**

**Model Architecture**

**Base Model:**

RoBERTa-base (a transformer-based model pre-trained on a large corpus of text).

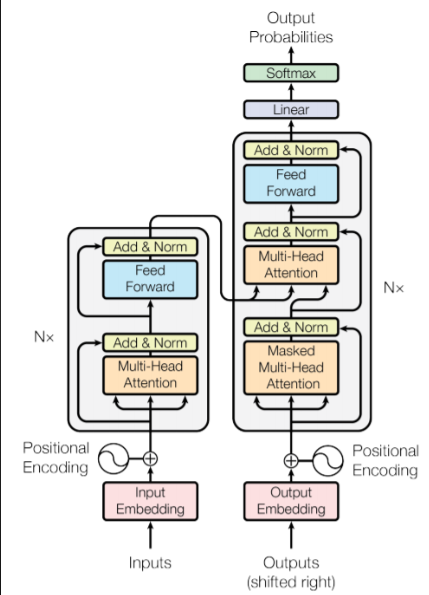


Then will be passed to a t5-base model to Alternate Phrasings

**Fine-Tuning:**

The model was fine-tuned for classification (`toxic` vs. `neutral`).

Added a classification head on top of the pre-trained RoBERTa model.



**Methodology**

**1. Data Preprocessing:**

**Dataset:** final\_labels.csv with columns body (comment text), level\_1 and split

**2. Tokenization:**

- Comments were tokenized using the RoBERTa tokenizer with a maximum sequence length of 128 tokens.

**3. Training:**

Fine-tuned the RoBERTa model using the TrainerAPI from Hugging Face.

It requires API key from the wandb

**Training parameters:**

* Batch size: 8
* Epochs: 3
* Learning rate: 2e-5
* Weight decay: 0.01

**4.Evaluation:**

The model was evaluated on the test set using accuracy, precision, recall, and F1 score.

**5.Alternative Phrasing:**

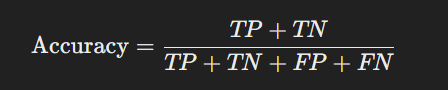
Detected toxic comments were passed to a T5-base model to generate alternative phrasings.

**3. Evaluation Metrics**

Metrics Used

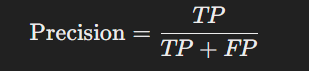
1. Accuracy:

- Measures the percentage of correctly classified comments.



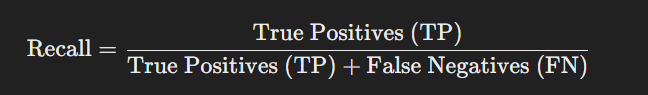
2. Precision:

- Measures the percentage of correctly predicted toxic comments out of all predicted toxic comments.

- 

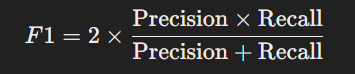
3. Recall:

- Measures the percentage of actual toxic comments correctly identified by the model.

- 

4. \*\*F1 Score\*\*:

- Harmonic mean of precision and recall.



-

**Results**

Accuracy: 90.08%

Precision: 87.02%

Recall: 90.08%

F1 Score: 87.29%

**4. Challenges Faced and Solutions Implemented**

Challenge 1: Imbalanced Dataset

- Problem:

- The dataset had more `neutral` comments than `toxic` comments, leading to biased predictions.

- Solution:

- Used class weights during training to give more importance to the minority class (`toxic`).

**5. Results and Examples**

Detected Toxic Comments

1. Original: "You're useless."

- Alternative: "You could improve in this area."

2. Original: "This is the worst idea ever."

- Alternative: "This idea could use some improvements."