Report on Fine-tuning T5 Model with Custom Dataset

Executive Summary

This report details the process and outcomes of fine-tuning a T5 model using a custom dataset. The task involved generating predictions for textual data, aiming to enhance the model's performance for specific use cases.

Introduction

The Transformer-based T5 (Text-to-Text Transfer Transformer) model has been widely adopted for various natural language processing tasks. Fine-tuning such a model on a domain-specific dataset can significantly improve its utility in specialized applications.

Methodology

Data Preparation

The data was preprocessed and prepared for fine-tuning as follows:
- The dataset was reordered because initial dataset reference and translation toxicity were not consistent
- The dataset was tokenized using the T5 tokenizer.
- A custom data collator was implemented to handle padding, ensuring consistent input batch sizes.

Model Training

- The model was trained using the Hugging Face Transformers library.
- A custom training loop was created, leveraging the Trainer API for efficient training.
- The model's performance wasn’t monitored properly because of its computational difficulty.

Prediction Generation

- A function was developed to load the fine-tuned model and generate predictions.
- This function was later modified to process single text strings and a limited number of entries (n) for predictions.
- Predictions were saved in TSV format for further analysis or deployment.