Cross-Domain Analysis for Detecting Human and AI-Generated Text

Divya Sree Kuntamalla, Saketh Reddy Dodda

CSCI 5922 – University of Colorado Boulder

Motivation

With the rise of large language models like ChatGPT, it has become increasingly difficult to distinguish between human-written and AI-generated content. While these models offer valuable tools for education, content creation, and customer service, they also introduce risks such as academic dishonesty, misinformation, and fake reviews. Most current detection systems are trained on limited benchmark datasets and are often designed for English text alone. This limits their reliability in real-world settings where language and topic diversity is common. Our project aims to address this gap by developing models that can accurately detect AI-generated text across multiple domains and languages.

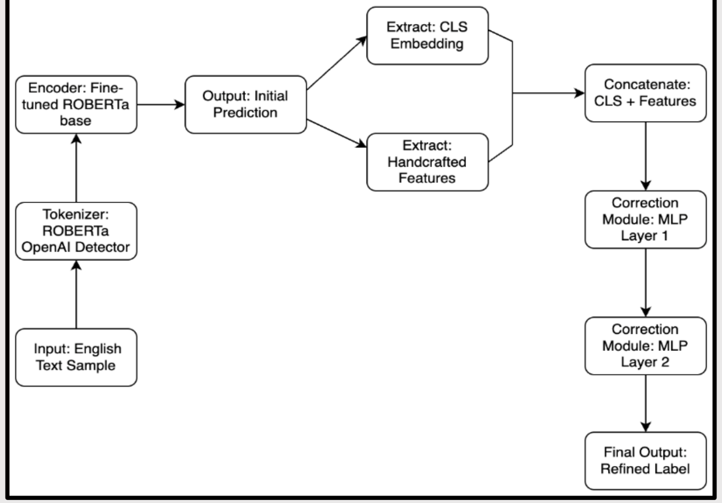
Why Existing Solutions Fall Short vs Our Approach

Existing models, trained on synthetic English-only datasets, often fail on real-world, multilingual content. Their performance drops when applied to new domains like product reviews or news articles, and especially when encountering unseen languages like French.

To overcome these limitations, we fine-tuned robust transformer models—RoBERTa for English (monolingual) and XLM-R for multiple languages (multilingual)—on benchmark datasets and tested them in cross-domain and cross-lingual environments. We evaluated their ability to generalize using real-world datasets like Book reviews & Amazon reviews, analyzing performance drop-offs and identifying strengths across settings.

Task	Dataset Type	Accuracy	Macro-F1
Monolingual	Validation (In-Domain)	**96.2%**	0.9619
	Test (Cross-Domain)	**62.4%**	0.6233
Multilingual	Validation (In-Domain)	**75.3%**	0.7512
	Test (Cross-Domain)	**53.1%**	0.5304

Monolingual Architecture



Input: English Text Sample

Tokenizer: ROBERTa OpenAI Detector

Encoder: Fine-tuned ROBERTa base

Output: Initial Prediction

Extract: CLS Embedding

Extract: Handcrafted Features

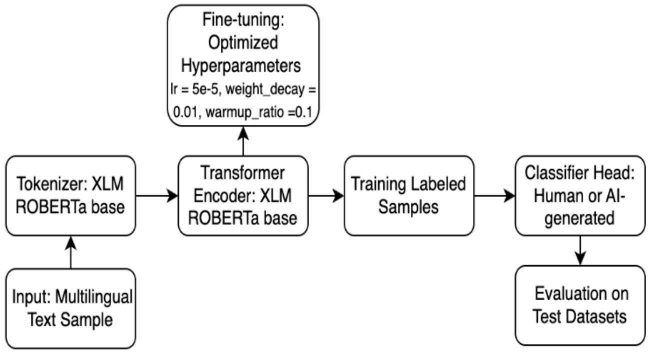
Concatenate: CLS + Features

Correction Module: MLP Layer 1

Correction Module: MLP Layer 2

Final Output: Refined Label

Multilingual Architecture



Input: Multilingual Text Sample

Tokenizer: XLM ROBERTa base

Transformer Encoder: XLM ROBERTa base

Fine-tuning: Optimized Hyperparameters lr = 5e-5, weight_decay = 0.01, warmup_ratio =0.1

Training Labeled Samples

Classifier Head: Human or AI-generated

Evaluation on Test Datasets