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

Divya Sree Kuntamalla, Saketh Reddy Dodda
Course CSCI 5922, University of Colorado Boulder

1 Introduction

With the increasing capabilities of large language models (LLMs), the distinction between human-authored and machine-generated text has become progressively less apparent. These models currently produce fluent, coherent, and contextually rich language over a wide range of concepts, making their outputs impossible to differentiate from those written by humans [2]. While such improvements provide significant benefits in domains such as content creation, education, and customer service, they also raise concerns about misinformation, academic dishonesty, and a loss of confidence in digital communication [3]. Accurately recognizing whether a piece of text was generated by a machine is thus an important task for digital platforms, educators, and content verifiers. The current research looks into the effectiveness of neural network-based algorithms for detecting machine-generated language, with a special emphasis on high-impact sectors like product reviews and journalistic articles.

Despite increased interest in AI-generated text detection, much of the current research has focused on synthetic benchmark datasets or general-purpose web text corpora [4]. These datasets, while beneficial for controlled research, may not capture the linguistic variety and contextual complexity present in application-specific domains like review platforms or news media. Furthermore, many detection systems are created and assessed only in English-language contexts, raising concerns about their resilience in multilingual environments. The existing literature provides little insight into the generalizability of pre-trained detection models across multiple content categories and languages. The lack of cross-domain and multilingual evaluation restricts the practical deployment of detection systems in real-world, globalized environments.

This project compares transformer-based models for distinguishing human-written versus AI-generated text in monolingual and multilingual contexts. The approach focuses on real-world domains, specifically book reviews, which contain both subjective and objective textual content. Monolingual evaluation involves fine-tuning pre-trained models like RoBERTa on English-language datasets, whereas multilingual evaluation uses XLM-RoBERTa to compare detection performance across many languages [1]. By applying these models to previously unexplored, domain-specific datasets, the study aims to evaluate their effectiveness in realistic situations and investigate their generalizability. The inclusion of monolingual and multilingual analysis broadens the perspective compared to past studies, demonstrating the strengths and limitations of current detection architectures in real-world applications.

2 Related Work

Topic 1 – Benchmark-Driven AI Text Detection

References:

- **Roy Dipta and Shahriar (2024).** HU at SemEval-2024 Task 8A: Can Contrastive Learning Learn Embeddings to Detect Machine-Generated Text? [4]

- **Aggarwal and Sachdeva (2024).** CUNLP at SemEval-2024 Task 8: Classify Human and AI Generated Text. [1]

While both HU and CUNLP submissions use pretrained models (such as transformer-based encoders), their approaches remain limited to competition-provided datasets and benchmark-specific evaluations. This project, by contrast, fine-tunes the same pretrained models on the same training data but shifts the focus toward evaluating their generalization ability in domain-specific contexts, such as book reviews and news articles. Rather than introducing new training resources, this work investigates how these models perform when applied to previously unseen content types, revealing limitations and strengths that are not apparent in in-domain evaluations. This cross-domain setup provides a more realistic perspective on the practical deployment of detection systems beyond competition environments.

Topic 2 – Deep Learning Architectures for Detecting Machine-Generated Text

References:

- **Guo et al. (2024).** DeTeCtive: Detecting AI-generated Text via Multi-Level Contrastive Learning. [2]

- **Latif et al. (2024).** Detection of AI-Written and Human-Written Text Using Deep Recurrent Neural Networks. [3]

In contrast to prior works, this project does not introduce novel neural architectures or training paradigms. Instead, it investigates the capabilities of existing, general-purpose transformer models in classifying AI-generated vs human-written text across multiple domains and languages. Both Latif et al. and Guo et al. emphasize model design over real-world deployment situations, whereas Guo et al. provide a contrastive learning methodology to improve model robustness. In the current project, popular pretrained transformers are applied to unknown, diverse, and naturally occurring datasets using a practical evaluation approach. This helps understand their effectiveness in more realistic detection scenarios where domain and linguistic variation is common.

3 Methodology

This project will investigate the use of transformer-based models to classify human-written versus machine-generated text in both monolingual and multilingual contexts. For the monolingual track, the model selected will be *roberta-base*, chosen for its proven effectiveness on English-language classification tasks and its computational efficiency. Fine-tuning will be performed using *LoRA* (Low-Rank Adaptation) to reduce the number of trainable parameters, allowing for faster experimentation and lower memory usage. Other lightweight fine-tuning strategies may also be explored depending on resource availability. For the multilingual track, *XLM-RoBERTa-base* will be used due to its strong performance across a wide range of languages. It will also be fine-tuned using *LoRA* or similar parameter-efficient methods. All models will be implemented using the Hugging Face Transformers library within a PyTorch framework and trained in GPU-enabled environments such as Google Colab. Input texts will be preprocessed using the appropriate tokenizer for each model and truncated or padded to a maximum sequence length of 512 tokens.

The training datasets used for both monolingual and multilingual settings will be drawn from a publicly available benchmark designed for binary classification of human and machine-generated text. These datasets will cover a variety of domains and languages. The key dataset statistics are as follows:

Monolingual Training Data:

- Machine-generated texts: 56,400
- Human-written texts: 63,351
- sources: Wikipedia, WikiHow, Reddit, arXiv, PeerRead

Split	Source	davinci-003	ChatGPT	Cohere	Dolly-v2	BLOOMz	GPT-4	Machine	Human
Train	Wikipedia	3,000	2,995	2,336	2,702	—	—	11,033	14,497
	Wikihow	3,000	3,000	3,000	3,000	—	—	12,000	15,499
	Reddit	3,000	3,000	3,000	3,000	—	—	12,000	15,500
	arXiv	2,999	3,000	3,000	3,000	—	—	11,999	15,498
	PeerRead	2,344	2,344	2,342	2,344	—	—	9,374	2,357
Dev	Wikipedia	—	—	—	—	500	—	500	500
	Wikihow	—	—	—	—	500	—	500	500
	Reddit	—	—	—	—	500	—	500	500
	arXiv	—	—	—	—	500	—	500	500
	PeerRead	—	—	—	—	500	—	500	500

Table 1: Monolingual dataset stats by data source and model. Training data was sourced from five domains, and the development set includes balanced BLOOMz and human written samples.

Multilingual Training Data:

- Machine-generated texts: 76,863
- Human-written texts: 80,994
- Languages: English, Chinese, Urdu, Bulgarian, Indonesian etc.,

Split	Language	davinci-003	ChatGPT	LLaMA2	Jais	Other	Machine	Human
Train	English	11,999	11,995	—	—	35,036	59,030	62,994
	Chinese	2,964	2,970	—	—	—	5,934	6,000
	Urdu	—	2,899	—	—	—	2,899	3,000
	Bulgarian	3,000	3,000	—	—	—	6,000	6,000
	Indonesian	—	3,000	—	—	—	3,000	3,000
Dev	Russian	500	500	—	—	—	1,000	1,000
	Arabic	—	500	—	—	—	500	500
	German	—	500	—	—	—	500	500

Table 2: Multilingual dataset stats across five training and three development languages.

Standard classification training techniques will be followed, with hyper-parameters such as batch size and learning rate selected based on validation performance. During training, accuracy and precision will be tracked as primary evaluation metrics, and confusion matrices will be used to visualize model behavior and error patterns. This approach will enable a consistent and reproducible comparison of model performance across different linguistic and domain-specific scenarios.

4 Experiments

Experiment 1: Multilingual Detection of Machine-Generated Text

Main Purpose: This experiment aims to evaluate the performance of the multilingual model (XLM-RoBERTa) on detecting machine generated text written in different languages. The objective is to test whether a model trained on benchmark multilingual data can generalize to real-world texts written in diverse languages and produced by modern language models.

Setup:

- Machine-generated texts will be created using ChatGPT, with approximately 200 samples per language generated through prompt-based review synthesis in the target languages.

- A smaller number of human-written reviews will be sampled from Kaggle's multilingual Amazon review datasets to create a realistic and unbalanced test set.
- Test data will combine these two sources per language to simulate a skewed real-world scenario.
- All texts will be preprocessed and tokenized using the appropriate multilingual tokenizer to match the training pipeline.

Evaluation Metrics:

- **Accuracy:** Measures overall correctness of predictions across languages.
- **Precision:** Ensures the model performs consistently across all language classes.
- **Per-language Confusion Matrix:** Highlights strengths or weaknesses in specific languages, useful for multilingual model analysis.

Experiment 2: Cross-Domain Evaluation on User-Generated Content

Main Purpose: This experiment will assess the domain generalization capabilities of the models trained on benchmark datasets by evaluating their performance on user-generated content, specifically in the context of book reviews. The goal is to determine whether the trained models can maintain performance when exposed to a new and practical content domain not seen during training.

Setup:

- Human-written book reviews will be collected from public datasets such as Amazon Book Reviews (via Kaggle).
- Machine-generated reviews will be created using ChatGPT and other open-source language models, prompted to produce realistic book reviews.
- Around 2,000 samples will be prepared, balanced between human and machine-written texts.
- All samples will be tokenized and structured consistently with the training data.

Evaluation Metrics:

- **Accuracy:** Indicates overall model performance.
- **Precision:** Critical for identifying how well the model avoids false positives—important when distinguishing subtle content like reviews.
- **Confusion Matrix:** Helps interpret types of classification errors made in a real-world domain like user-generated reviews.

References

[1] Pranjal Aggarwal and Deepanshu Sachdeva. CUNLP at SemEval-2024 task 8: Classify human and AI generated text. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1–6, Mexico City, Mexico, 2024. Association for Computational Linguistics.

[2] Xun Guo, Yongxin He, Shan Zhang, Ting Zhang, Wanquan Feng, Haibin Huang, and Chongyang Ma. Detective: Detecting ai-generated text via multi-level contrastive learning. *Advances in Neural Information Processing Systems*, 37:88320–88347, 2024.

[3] Ghazanfar Latif, Nazeeruddin Mohammad, Ghassen Ben Brahim, Jaafar Alghazo, and Khaled Fawagreh. Detection of ai-written and human-written text using deep recurrent neural networks. In *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)*, volume 13162, pages 11–20. SPIE, 2024.

[4] Shubhashis Roy Dipta and Sadat Shahriar. HU at SemEval-2024 task 8A: Can contrastive learning learn embeddings to detect machine-generated text? In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 485–491, Mexico City, Mexico, 2024. Association for Computational Linguistics.