Proposal: Detecting Machine-Generated Text Across Domains and Languages

Title :

Bot or Not? : Transformer-Based Machine-Generated Text Detection Across Domains

Team Details

- **Name:** Mohammed Bouadjimi
- **Student ID:** 041179141
- **Course:** CST8507 – Natural Language Processing
- **Term:** Spring/Summer 2025
- **Institution:** Algonquin College

Topic and Motivation

For this assignment, I chose to work on **SemEval-2024 Task 8: Machine-Generated Text Detection**, a challenge that focuses on distinguishing between human-written and AI-generated texts. With the rise of advanced language models like ChatGPT and GPT-4, it's becoming harder to differentiate AI-generated content from human writing. This poses ethical, academic, and security concerns—ranging from misinformation to academic dishonesty.

This task is not only timely but personally interesting, especially as someone who is studying NLP. I wanted to explore how well machines can detect other machines, using the same kinds of models that generate the text in the first place.

Dataset and Tools

I will use the official dataset provided for SemEval-2024 Task 8. This dataset consists of English texts labeled as either human-written or machine-generated, pulled from various domains including academic writing, news articles, and Reddit discussions.

- **Data Source:** Official SemEval 2024 repository
- **Data Format:** JSON/CSV with text and label fields
- **Labels:** `human` or `machine`
- **Preprocessing Tools:** NLTK, spaCy, HuggingFace Datasets
- **Modeling Tools:** HuggingFace Transformers (specifically BERT or RoBERTa), PyTorch/Colab

---

Plan of Work

Milestone	Estimated Date
Dataset loading and preprocessing	July 1–2
Baseline model training (Logistic/MLP)	July 4–6
Transformer model fine-tuning (BERT)	July 7–10
Evaluation (F1-score, ROC, confusion)	July 11–14
Final tweaks + error analysis	July 15–18
Report writing and formatting	July 19–26
Final submission + presentation prep	July 27–28

Expected Results and Interest

I hope to fine-tune a transformer model (like RoBERTa) that achieves a **high F1-score** on detecting machine-generated text, especially across different domains. I expect that:

- RoBERTa will outperform baseline models like Logistic Regression.
- The task will expose interesting limitations in both models and detection approaches.

This problem matters because **machine-generated text detection is now critical in many real-world settings**—from education to journalism to online platforms. Solving this problem means understanding how AI imitates human language—and how we can spot the difference.