Title: Different Learning Methods for AI-Generated Text Detection

Author: Ng Man Tik

Introduction (Problem Statement): The ubiquity of generative models like ChatGPT and GPT-4 has led to increased use of AI in fields such as academic writing and creative content generation. While beneficial, this raises issues regarding authenticity and potential misuse, as users might submit AI-generated texts as their own work. This project addresses the challenge of distinguishing between AI-generated and human-written texts, focusing on various contexts including text length, model capabilities, and mixed content. The relevance of this investigation stems from the growing overlap between AI-generated and human-produced texts, demanding more sophisticated detection methods.

Dataset (Data Set): The study will utilize the M4 dataset [11], a comprehensive collection featuring texts from multiple domains and languages, specifically focusing on English content. This dataset includes 122k entries from both AI-generated and human sources. For testing, the Mixset dataset, comprising 3.6k mixed instances across different formats, will be used to evaluate the detection of blended content [3]. Additionally, the MGTBench dataset will provide 2.4k samples of Human and AI-generated essays, stories, and news articles for model testing, addressing the detection capabilities across varied text types and lengths [12].

Dataset	Source	Size	Part Used	Potential Obstacles
M4	Wikipedia, Reddit ELI5, WikiHow, PeerRead, arXiv	122,481 entries	English texts and certain field of source only	High diversity of source will affect the consistency and model training
MixSet	Real-life mixed cases of AI-generated and human-edited texts in email, game review and total 6 categories	3,600 entries	Use all of them in Current Plan	Models training with purely AI-text may not detect well on mixed text
MGTBench	Human and 7 different AI models' generated text in Essay, Story writing and articles	24,000 entries	Currently planning to only adapt the data generated by GPT4ALL only, but may use other data for training process.	The quality and the length of the data varies.

Methodology (Methods):

This research will explore and enhance various deep learning techniques for detecting AI-generated text:

1. **Hybrid CNN-RNN Models:** The project will implement a text classification model combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), specifically Long Short-term Memory (LSTM) networks. This approach is designed to exploit CNNs for local feature extraction and RNNs for capturing the sequential nature of text. The project will try to adapt the model from [6] first. The improvement will include experimenting with different layer architectures and hyperparameters to optimize performance for varying text lengths and complexities as well as integrating the TF-IDF for context-based vectorization.
2. **Support Vector Machines (SVMs):** SVMs with Radial Basis Function (RBF) kernels will be employed for baseline text classification. The improvement will involve integrating advanced text feature extraction techniques, such as word embeddings or context-based vectorization, beyond the traditional TF-IDF approach. This modification aims to enhance the SVM's ability to understand semantic relationships within the text, potentially increasing the accuracy of AI text detection. Meanwhile trying to replace RBF with other kernel functions (Polynomial, Linear) to test their performance in text classification following the paper [5].
3. **Zero-Shot Learning:** Utilizing the OpenAI API, this approach will leverage ChatGPT-3.5 for zero-shot learning text classification. The improvement here will focus on refining the prompt engineering to better align with the detection task and exploring the impact of different instruction formats on classification accuracy.
4. **Fine-Tuning Language Models:** The project plans to fine-tune smaller-scale language models, such as miniCPM 2B [9], utilizing bf16 training, DeepSpeed, and QLoRa to enhance computational efficiency and reduce training time. The modifications will include customizing the training dataset to better reflect the mixed nature of real-world texts and adjusting the training regime to focus on distinguishing subtle differences between AI-generated and human-written content. The project will first utilize the original model from HuggingFace before fine-tuning. Moreover, if resources permit, we will experiment with fine-tuning alternative models, such as RoBERTa, to compare performance and understand the trade-offs between model size and detection accuracy.

Each method will be systematically evaluated and refined based on performance

metrics and detection effectiveness in real-world text scenarios. The ultimate goal is to develop an optimized approach that can accurately differentiate between human and AI-generated texts, even in mixed or nuanced cases.

Evaluation and Expected Results:

The effectiveness of the developed models will be evaluated using a multi-faceted approach:

1. **Quantitative Evaluation:** The primary metrics for quantitative assessment will be Area Under the Curve (AUC) and F1 scores, which provide a balanced view of model precision and recall. We will compare these metrics across different models to evaluate their performance in detecting AI-generated texts under various conditions (e.g., text length, complexity, and AI vs. human origin). The hypothesis is that hybrid CNN-RNN models and fine-tuned LLMs will outperform traditional SVMs and fine-tuned LLMs outperform zero-shot LLMs, especially in complex text scenarios and mixed AI-human content.
2. **Qualitative Evaluation:** The models' qualitative performance will be assessed through field analysis and case studies of specific detection instances. This will involve examining the performance of the models in different fields in the dataset and identifying patterns or characteristics of texts that are commonly misclassified. Additionally, attention maps and other interpretability techniques will be used to understand the decision-making process of neural models.
3. **Comparative Analysis:** The results will be compared against baseline models, such as zero-shot learning with ChatGPT-3.5 and traditional SVM classifiers. The hypothesis is that tailored approaches, especially those involving fine-tuning on relevant datasets, will provide superior detection capabilities compared to generic or pre-trained models.

Reference List (Including the paper / websites read to get the ideas)

[1] “LLM - Detect AI Generated Text,” Kaggle, 2024. [Online]. Available: https://www.kaggle.com/competitions/llm-detect-ai-generated-text3

[2] B. Priyamvada, S. Singhal, A. Nayyar, R. Jain, P. Goel, M. Rani, and M. Srivastava, “Stacked CNN - LSTM approach for prediction of suicidal ideation on social media,” in Multimedia Tools and Applications, vol. 82, pp. 27883–27904, Feb. 2023. [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-14431-z1

[3] C. Gao, D. Chen, Q. Zhang, Y. Huang, Y. Wan, and L. Sun, “LLM-as-a-Coauthor: The Challenges of Detecting LLM-Human Mixcase,” in arXiv:2401.05952, Jan. 2024. [Online]. Available: https://arxiv.org/abs/2401.059522

[4] D. Chen, “MixSet: Official code repository for Mixset,” GitHub, 2024. [Online]. Available: https://github.com/Dongping-Chen/MixSet1

[5] D. Prastyo and S. Ardiyanto, “Indonesian Sentiment Analysis: An Experimental Study of Machine Learning and Deep Learning Methods,” Semantic Scholar, Oct. 2020. [Online]. Available: https://www.semanticscholar.org/paper/Indonesian-Sentiment-Analysis%3A-An-Experimental-of-Prastyo-Ardiyanto/8a55cea7e1e12b599e65ec25205330916b4f706c

[6] J. A. Nasir, O. S. Khan, and I. Varlamis, “Fake news detection: A hybrid CNN-RNN based deep learning approach,” in International Journal of Information Management Data Insights, vol. 1, no. 1, 100007, 2021. [Online]. Available: https://doi.org/10.1016/j.jjimei.2020.100007

[7] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, and L. S. Chao, “A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions,” in arXiv:2310.14724, Oct. 2023. [Online]. Available: https://arxiv.org/abs/2310.147241

[8] K. Wang, “Fake News Detection,” GitHub, 2024. [Online]. Available: https://github.com/karenyxwang/Fake_News_Detection2

[9] OpenBMB, “MiniCPM: An end-side LLM outperforms Llama2-13B,” GitHub, 2024. [Online]. Available: https://github.com/OpenBMB/MiniCPM1

[10] S. Lyu and J. Liu, “Combine Convolution with Recurrent Networks for Text Classification,” in arXiv:2006.15795v1, Jun. 2020. [Online]. Available: https://arxiv.org/abs/2006.15795v11

[11] X. He, X. Shen, Z. Chen, M. Backes, and Y. Zhang, “MGTBench: Benchmarking Machine-Generated Text Detection,” in arXiv:2303.14822, Mar. 2023. [Online]. Available: https://arxiv.org/abs/2303.148222

[12] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, C. Whitehouse, O. M. Afzal, T. Mahmoud, A. F. Aji, and P. Nakov, “M4: Multi-generator, Multi-domain, and Multi-lingual Black-Box Machine-Generated Text Detection,” in arXiv:2305.14902, May 2023. [Online]. Available: https://arxiv.org/abs/2305.149024