Literature Review - LLM Generated Text Detection

Abde Manaaf Ghadiali - G29583342, Gehna Ahuja - G35741419, Sagar Sheth - G32921700, Venkatesh Shanmugam - G27887303

Accurately detecting LLM-generated text (LGT) is becoming increasingly critical as AI text generation models continue to evolve. Several studies emphasize the importance of high-quality and sufficient data for training effective LGT detection models. Research by [[1]](#Ref1) demonstrates that an extensive token length (word count) is crucial for accurate classification. [[2]](#Ref2), [[3]](#Ref3) and [[4]](#Ref4) further emphasize the need for diverse training data encompassing various writing styles and sources, such as social media, scientific publications, creative writing (stories, essays, poetries) and computer programs (Python scripts). This diversity helps models learn the nuances of human language across different contexts.

Deep learning approaches, Random-forest, Support Vector Machine (SVM), Long Short-Term Memory (LSTM) and particularly transformer architectures like RoBERTa, show promising results in LGT vs human-generated text (HGT) detection. Studies by [[2]](#Ref2) and [[4]](#Ref4) highlight the ability of these models to capture subtle language features and identify patterns within text data. This allows them to distinguish between the statistical regularities inherent in machine-generated text and the natural variations found in human-written language.

As AI text generation models become more sophisticated, LGT detection methods need to adapt. Research by [[3]](#Ref3) underlines the importance of generalizability, interpretability, and resilience against adversarial manipulation. Generalizable models can effectively detect LGT across various domains and writing styles. Interpretable models allow us to understand how the models make their decisions, which is crucial for building trust in their accuracy. Finally, robust detection methods should be resilient against adversarial techniques employed to evade detection, ensuring the integrity of the LGT detection system.

The study by [[5]](#Ref5) highlights the need for tailored LGT detection methods for specific domains, such as the medical field. Their research demonstrates that linguistic analysis, alongside machine learning models, is crucial for accurate detection in specialized domains where precise language usage is necessary. Medical documents often require a specific vocabulary and sentence structure that can differ from more general writing styles. By incorporating domain-specific knowledge, LGT detection methods can achieve greater accuracy in these specialized contexts.

By expanding on prior research, we aim to enhance the accuracy and effectiveness of methods which classify between HGT and LGT.

**References**

**[[1] G. Gritsay, A. Grabovoy and Y. Chekhovich, "Automatic Detection of Machine Generated Texts: Need More Tokens," 2022 Ivannikov Memorial Workshop (IVMEM), Moscow, Russian Federation, 2022, pp. 20-26, doi: 10.1109/IVMEM57067.2022.9983964.](https://ieeexplore.ieee.org/document/9983964)**

**[[2] Gaggar, Raghav, Ashish Bhagchandani, and Harsh Oza. "Machine-Generated Text Detection using Deep Learning." arXiv preprint arXiv:2311.15425 (2023).](https://arxiv.org/abs/2311.15425)**

**[[3] Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. "Automatic detection of machine generated text: A critical survey." arXiv preprint arXiv:2011.01314 (2020).](https://arxiv.org/abs/2011.01314)**

**[[4] Kadhim Hayawi, Sakib Shahriar, Sujith Samuel Mathew," The Imitation Game: Detecting Human and AI-Generated Texts in the Era of ChatGPT and BARD",](https://arxiv.org/ftp/arxiv/papers/2307/2307.12166.pdf)**[**doi.org/10.1177/016555152412275**](http://doi.org/10.1177/016555152412275)[**.**](https://arxiv.org/ftp/arxiv/papers/2307/2307.12166.pdf)

**[[5] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, Xiang Li, "Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study", doi: 10.2196/48904](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10784984/)**