PROJECT REPORT

DE-GPT: A MODEL DETECTOR TO DISTINGUISH BETWEEN LLMS AND HUMAN TEXT

ALEXANDRO SETIAWAN
19.K1.0037

Faculty of Computer Science
Soegijapranata Catholic University
2023

ABSTRACT

The advent of large language models (LLMs), including ChatGPT Google Bard, and Bing AI, has had a transformative impact on natural language processing. However, the rise of these models has also introduced a new challenge: distinguishing between text produced by humans and that generated by LLMs. Accurate detection of LLM-generated text is vital due to the risk of disinformation, fake news, and automated spam, and has significant implications across various fields such as journalism and social media. This study aims to tackle this challenge by fine-tuning the GPT-Neo model with variant of 1.3B parameters to detect and identify text generated by LLMs or text produced by human. The objectives include developing a fine-tuned GPT-Neo 1.3b model for generated text detection, evaluating its performance, and conducting a comparative analysis with other existing models. This research will utilize a customized dataset comprising both LLM-generated and human-authored texts. The results of this research could provide valuable insights into the practical applications and potential implications of accurate text generation detection in real-world settings.

Keyword: large language models, ChatGPT, fake text detection, GPT-Neo