AI Content Detector

Synopsis

A Synopsis Submitted

in Partial Fulfillment of the Requirements

for the Degree of

**BACHELOR OF TECHNOLOGY**

in

# Department of Artificial Intelligence and Artificial Intelligence & Machine Learning

by

# Devashree Bhosale A19

# Eshan Kasliwal A22

# Devang Pagare A18

# Pratik Jade A70

****

**G H Raisoni College of Engineering and Management**

**Wagholi, Pune**

**Department of Artificial Intelligence and Artificial Intelligence & Machine Learning**

**(An Autonomous Institute Affiliated to SPPU, Pune)**

# August 2023

1. **Introduction**

The rise of artificial intelligence (AI) technology and its impact on education have been topics of growing concern in recent years. The new generation of AI systems, such as chatbots, have become more accessible on the Internet and stronger in terms of capabilities. The use of chatbots, particularly ChatGPT, for generating academic essays at schools and colleges has sparked fears among scholars. This study aims to explore the originality of the content produced by one of the most popular AI chatbots, ChatGPT.

The AI Content Detector project aims to develop an advanced system that utilizes artificial intelligence and machine learning techniques to automatically detect and classify different types of content written by AI. AI detectors may be used to detect when a piece of writing is likely to have been generated by AI. This is useful, for example, to educators who want to check that their students are doing their own writing or moderators trying to remove fake product reviews and other spam content.

ChatGPT is a conversational artificial intelligence that is a member of the generative pre-trained transformer family of the large language model family. This text-generative model was finetuned by both supervised learning and reinforcement learning so that it can produce text documents that seem to be written by natural intelligence. Although this generative model has numerous advantages, it also comes with some reasonable concerns. This project presents a machine learning-based solution that can identify the ChatGPT-delivered text from the human-written text, along with a comparative analysis of algorithms in the classification process.

1. **Motivation**

The rapid advancement of artificial intelligence has led to a growing need for tools to identify the source of content. By identifying the distinctive characteristics of AI and human-generated content, this project aims to provide a tool for identifying misinformation, authenticity, intellectual property protection, responsible AI use, user empowerment, bridging the gap between AI and humans, ethical AI development, and meeting societal needs.

AI content detectors can be used by journalists to verify the authenticity of news articles, by social media users to detect fake news and propaganda, and by teachers to identify students who are using AI-generated text. This not only preserves the integrity of their work but also safeguards their competitive advantage in the market. By detecting instances of plagiarism and copyright infringement, businesses can also take legal action to protect their rights and reputation.

1. **Related work**

To detect AI-generated texts multiple approaches are proposed. Large language models (LLMs) are statistical language models trained to predict the likelihood of a specific character, word, or string (called a token) in a particular context. LLMs have been used since the 1980s for various tasks, including machine translation and automatic speech recognition.

Early LLMs were trained using a technique called n-gram language modelling. In n-gram language modelling, the probability of a token is estimated based on the n-grams that have come before it. For example, the probability of the token "the" is estimated based on the two grams that have come before it, such as "in the" and "the cat".

In the 2010s, new methods for training LLMs were developed. These methods, such as word embedding and transformer architecture, made it possible to train LLMs that are much larger and more powerful than previous models. These LLMs are able to generate human-like text, translate languages, and answer questions in an informative way.

ChatGPT is a Natural Language Processing system that is owned and developed by OpenAI, a research and development company established in 2015. Based on the transformer architecture, OpenAI released the first version of GPT in June 2018. Within less than a year, this version was replaced by a much improved GPT-2, and then in 2020 by GPT-3. (Marr, 2023). This version could generate coherent text within a given context. This was in many ways a game-changer, as it is capable of creating responses that are hard to distinguish from human-written text.

The development of LLMs has led to an acceleration of different types of efforts in the field of automatic detection of AI-generated text. Firstly, several researchers have studied human abilities to detect machine-generated texts. like (DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature ). Secondly, (Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning). Thirdly, (CHATGPT OR HUMAN? DETECT AND EXPLAIN.) These efforts provide a valuable contribution to improving the performance and capabilities of detection tools for AI-generated text

1. **Proposed Method**

The AI Text Content Detector project employs a comprehensive methodology to develop an effective system for analyzing and classifying text content. The proposed method encompasses data collection, preprocessing, model selection, training, evaluation, real-time monitoring, user interface design, feedback integration, and continuous adaptation.

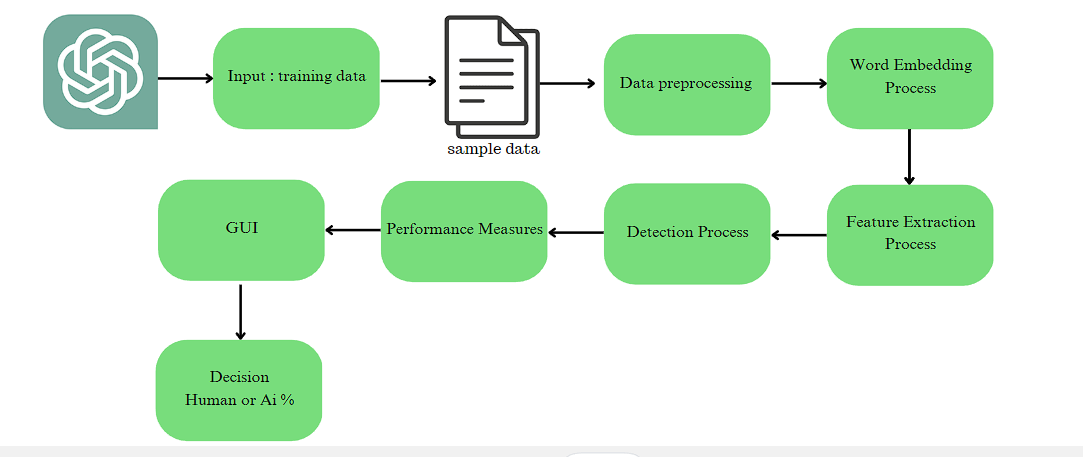
****

figure 1

1. **Methodology**

The aim of this project is to differentiate human text from generative model text using machine learning. In Figure 1, the high-level overview of our process is described. The task is initiated by data collection

Our approach consists of two main building blocks. The first consists of a machine learning model trained to discriminate between text samples generated by a human and text samples generated by ChatGPT. The second is an explainable AI framework to explain and interpret the outcomes of the ML model used for discrimination. We need to understand the inner workings of the obtained model aiming to derive the insights about writing styles of humans, on the one hand, and that of ChatGPT, on the other, and their subsequent differences.

Different directions could be pursued in order to extract useful features from a piece of text and perform text classification. Given the recent success of Transformer-based architectures [3, 13] in solving different NLP tasks, we decided to follow the same vein to build our classifier. As such, we start with a Transformer-based model pre-trained for the sequence classification task, using its corresponding tokenizer to preprocess data. We then fine-tune the model on the training subset of collected data to detect whether the text sample belongs to the positive (ChatGPT-generated) or the negative (human-generated) class. Consequently, the fine-tuned model is used for inference on the testing subset. Finally, the obtained classification scores are evaluated against the ground truth.

1. **Plan of work**

* August: Data Collection and Model Selection
* September: Model Development and Evaluation
* October: Deployment and User Interface
* November: Final Review and Refinement

****

**References**

1. https://arxiv.org/pdf/2301.13852.pdf
2. https://arxiv.org/pdf/2306.01761.pdf
3. https://arxiv.org/pdf/2301.11305v1.pdf
4. https://arxiv.org/ftp/arxiv/papers/2302/2302.04335.pdf
5. https://ieeexplore.ieee.org/document/10137767
6. https://ieeexplore.ieee.org/document/9983964
7. https://www.zerogpt.com/
8. https://detectgpt.ericmitchell.ai/
9. https://github.com/jhroy/auditGPTZero
10. https://huggingface.co/datasets/artem9k/ai-text-detection-pile

**Dr. J. Y. Deshmukh Prof. R. Y. Sable**

**Project Guide Project Coordinator HOD**