

## **Individual Project Topic**

**Title:** Comparative Analysis of AI-Generated and Human-Written Essays

**Name:** Weiqian peng

**Date:** 6/17/2024

## **Introduction**

In the digital age, artificial intelligence (AI) has become a key force in various sectors, revolutionizing the field of content creation in particular. With the development of AI technologies, they are increasingly being used to generate written content ranging from simple reports to complex narrative essays. The rapid development trend and increasing use of AI compared to human-written text makes the quality of the content it generates an important challenge to consider. The project aims to delve into qualitative differences between papers written by AI algorithms and those written by human authors. By analyzing a massive dataset containing about 500,000 articles - evenly divided between AI-generated and human-written texts - the study aims to reveal stylistic, thematic, and structural differences between the two sources. As educational institutions and content-dependent industries consider incorporating AI writing tools into their operations, understanding these differences will highlight the current capabilities and limitations of AI in mimicking human creativity and coherence. In addition, the study takes place against the backdrop of an ongoing debate about the veracity and ethical implications of AI in authorship. By establishing a rigorous analytical framework to examine the nuances of AI and human papers, the project will provide empirical evidence to inform this debate, providing insights into how AI can be tailored to improve its ability to mimic and generate.

## **Research Question**

The central question of this study is "How do AI-generated articles differ from human-written articles in terms of style, complexity, and attractiveness to readers? This question is significant because it explores the capabilities and limitations of AI in mimicking human writing, a fundamental aspect of intelligence and communication. Previous research has focused on detecting AI-generated texts or assessing their readability, but few studies have conducted extensive comparative analyses using advanced text analysis techniques. In this project I will use novel methods of analysis, such as the use of linguistic tools and visual analysis techniques, to provide a comprehensive analysis of textual differences.

## **Materials/Methods**

To conduct this research, I will utilize a robust dataset comprising approximately 500,000 essays. This dataset is uniquely structured from Kaggle, containing an equal split between texts generated by artificial intelligence and those written by human authors. This diverse collection will allow for a comprehensive comparison across various dimensions of textual analysis.

Pandas: This library will serve as the backbone for data manipulation tasks. It is particularly well-suited for handling large datasets like ours, enabling efficient sorting, filtering, and

grouping of data. Pandas will also be used to prepare the dataset for further analysis, including cleaning and normalizing the data where necessary.

NLTK (Natural Language Toolkit): Essential for the natural language processing (NLP) aspects of the project, NLTK will be used to analyze the linguistic structures of the essays. This includes tokenization, part-of-speech tagging, and extraction of syntactic patterns.

Seaborn and Matplotlib: These visualization libraries will be employed to create compelling graphical representations of the data. Seaborn will be particularly useful for generating high-level statistical graphics in an intuitive manner, while Matplotlib will allow for more detailed customization of plots. Together, these tools will enable the visualization of complex relationships in the data, such as differences in syntax complexity and sentiment between the two sets of essays.

Techniques: The project will employ a series of statistical tests to quantitatively compare the linguistic characteristics of articles. These tests will help determine if there are statistically significant differences in syntactic complexity, word frequency distribution, and overall emotion between AI-generated and human-written texts. Syntactic complexity will be assessed by examining sentence structure and the use of various grammatical components. I will analyze the differences between humans and artificial intelligence in constructing complex sentences. I will look closely at word frequency distribution to identify patterns in word use, including the prevalence of certain words or phrases that may distinguish AI from human writing. I will also conduct sentiment analysis to measure the emotional tone of the article, comparing how AI and humans express positive, negative, or neutral emotions.

## **Results**

I expect that AI-generated articles and human-written articles will be very different in terms of stylistic elements and content engagement. One hypothesis is that AI-written articles may exhibit overly standardized or unusually complex patterns, fixed patterns in the content of sentences, or even repetitive ideas, because the learning algorithm is trained on different literature but at the same time lacks innovative thinking patterns. Another hypothesis is that human essays would show greater diversity in style and creative expression. The expected results will include a detailed comparison of these aspects and a visual approach to highlight the salient features and commonalities of the two types of articles.

## **Discussion**

Through this research, I hope to gain insight into the strengths or weaknesses of AI in replicating human writing. This will help further understand the potential of AI in educational Settings, such as automated grading or assisting in creative writing courses. The results may shed light on how to more accurately distinguish between AI and human writing. This research contributes to the academic discussion and optimization of the ability of artificial intelligence in text generation, and also provides practical implications for the future application of artificial intelligence in various fields.