AI vs Human Essays: A Data Visualization Project

Rida Arif

# Project Overview

## Hypothesis/Research Scope

Humans and AI use language differently, not just in structure but in purpose. AI-generated essays are more commonly used for certain topics that may require more research, whereas humans tend to write about subjects that they may already have prior knowledge of. Additionally, AI and human-written texts exhibit distinct word usage patterns—AI-generated text may lean toward more formal or neutral language, while human writing may show more subjective expression.

This project examines linguistic patterns in school essays written by humans versus those generated by AI, aiming to identify differences in word usage and topic distribution through sentiment analysis. The dataset, “Augmented data for LLM” by Johnathan Herrera, sourced from [Kaggle](https://www.kaggle.com/datasets/jdragonxherrera/augmented-data-for-llm-detect-ai-generated-text), consists of labeled essays—0 for human-written and 1 for AI-generated. Data preprocessing, feature extraction, and various visualization techniques were employed to get insights into how AI-generated content differs from human writing. If this were a full blown research project, my tentative research question may be “Do AI and human-written essays exhibit distinct differences in topic distribution and word usage patterns?”

## Summary Statistics

Of the 40,000 sampled essays, approximately 14674 (36.69%) were AI-generated, and 25325 (63.31%) were human-written. The average essay length was 387.20 words, with the shortest being 42 words and the longest 1368 words. Sampling the dataset may have skewed the results, as there are more human-written essays in this sample.

# Data Management

The dataset was first loaded into a pandas DataFrame for processing. Text data was standardized by converting all text to lowercase and removing special characters, punctuation, and numbers to ensure a consistent analysis. Stopwords were filtered out using spaCy's built-in list, and lemmatization was applied to reduce words to their base forms, aiding in meaningful comparison between AI and human-generated text.

The dataset has over 80,000 observations, of which half were sampled in this project due to device memory constraints. Due to the large size of even half the dataset, the project was mainly processed and written on Google Colab due to their GUI feature. The cleaning steps ensured that irrelevant or extraneous text did not affect word frequency or topic modeling results. Additionally, essays were tokenized, and part-of-speech (POS) tagging was conducted to extract lexical features for further analysis.

# Methodology Explanation

The analysis involved a combination of linguistic processing, topic modeling, and visualization techniques. The project’s original aim and hypothesis was a bit different than what the final version looks like today. The original hypothesis was that AI-generated essays may tend to use more predictable and formulaic phrasing whereas human-written essays exhibit more varied sentence structures and language, hence the focus was more on identifying the phrasing differences via trigram analysis. However, due to some limitations that will be further discussed in the **critical analysis** section, the focus was pivoted to center more on topic modelling.

## Visualization overview

A treemap and bar graphs were chosen to display topic distributions effectively, as they allow for categorical and hierarchical organization of key themes in AI and human-generated essays. Latent Dirichlet Allocation (LDA) was initially used to discover topics within the essays, but because LDA sometimes grouped words incorrectly and misrepresented the topics, a keyword-based topic mapping approach was introduced to refine topic classification. This ensured that words were assigned to relevant themes more accurately.

Word clouds were employed to visually represent the most frequently used words in AI and human-written essays, providing an intuitive overview of lexical differences. By leveraging TF-IDF and CountVectorizer, key words and phrases were extracted to further distinguish AI-generated essays from human-written ones.

### Visualization 1: Vertical bar graph – What do Humans Write About vs. What They Use AI For

A graph of different colored bars

AI-generated content may be incorrect.

This visualization was based on the categorization of AI and Human essays and the analysis of their topic distribution to see if there are any differences in trends. Based on my keyword-based topic mapping approach, I assumed that humans might resort to AI to write essays on topics that would require more research and on average would not have that much general knowledge on (e.g. Space Exploration, Automotive Vehicles). While on the other hand, there may be more human-written essays on things that affect people on a day-to-day basis, such as Motivation, Education and Politics.

As seen in the bar graph, the result was a bit of a mixed bag. Indeed, there were more human-written essays on Motivation like I had predicted, but I was surprised that Technology had an equal number of essays written by AI and humans. AI overwhelmingly dominated all the other essay topics, to my surprise, as I expected at least Education to be written more by humans. However, this visualization does not necessarily mean that humans resort to AI for all essays, and the topic categorization is also not 100% accurate, so serious conclusions must not be drawn from this visualization or data.

While this visualization is simple to look at, it does effectively depict what I aimed for it to do. I tinkered with the color scheme and typography quite a bit, ensuring easy readability and colors that were easily discernable from each other, but still pleasing to the eye. Though simple, it does have a toggleable legend and the ability to hover over each bar and see the exact number of essays.

### Visualization 2: Treemap – What do Humans Write About vs. What They Use AI For

A screenshot of a computer screen

AI-generated content may be incorrect.

Since I will have another bar graph (though completely different) later on, I thought it would be interesting to visualize this information in a different way. I chose a treemap as it’s an effective visual tool that presents data in a comparative and engaging way. By dividing the chart into two sections—topics related to human writing and those associated with AI usage—it allows for an intuitive visual comparison. I liked the proportional sizing of the rectangles, and with this gradient color scheme from blue to red, it makes it easy to identify the significance of each topic at a glance. The color scheme took quite a bit of troubleshooting, but landing on this slightly pastel red-blue seemed like a good landing point. I chose to include percentage values within the rectangles provides precise insights and allow viewers to quickly interpret the data without needing to think about the raw numbers too much.

However, the visualization could have been enhanced further, especially in regards to the color scheme. I did consider that the gradient may cause confusion when comparing the two sides of the treemap, and maybe a unified gradient for each AI and Human, such as different shades of blue for AI and red for humans, could make the comparisons more straightforward. I did try doing that, but after a number of failed attempts (the colors would constantly get mixed up), I settled on weighting the colors based on the number of essays throughout the treemap, whether or not they are AI or Human.

### Visualization 3: Word Clouds (can be viewed in the dashboard)

I wanted to also include word clouds as visualizations of this data, as I thought it would be interesting to visually see, specifically based on topics, which words are used the most. While coding these visualizations is quite straightforward, it is this step that made me realize that solely using LDA and TF-IDF as categorization methods was sometimes putting words into the wrong categories (e.g. “mars” in the motivation category). That is why I chose some specific keywords per category for topic mapping, to further refine the categories. It took a bit of trial and error, but in the end, the word clouds seemed to be representative, though the preprocessing did cut off some words (you might see “venu” instead of “venus”) but that seems like a minor issue. The word clouds can be viewed in the dashboard as including all of them would clutter up this document.

### Visualization 4: AI vs. Human Word Usage Difference

A graph with red and blue bars

AI-generated content may be incorrect.

For this visualization, I chose words that I thought were meaningful in the context of AI-written and human-written work in my experience. For instance, in essays AI tends to use formal, analytical words (Significant, Conclusion, Analysis) whereas human students may use more personalized words (Feel, Opinion, Believe). I wanted to depict this through a representation of a frequency difference, where:  
word\_freq\_df["freq\_diff"] = word\_freq\_df["human\_freq"] - word\_freq\_df["ai\_freq"]

Hence, the scale ranges from -4000 to 8000, as we are seeing the difference in each word’s frequency from its usage in human essays.

Initially, the code above was flipped so we were subtracting the *human* word frequency from the AI frequency instead, so the x-axis was showing a range from -8000 to 2000. After getting the opinions of laypeople (my flatmates), they thought the x-axis favoring negative values was confusing, leading to what the visualization looks like today.

Overall, I think this graph does a relatively good job at visually comparing word frequency differences between AI-generated and human-written essays. I chose a horizontal bar graph for this because it allows for easy side-by-side comparison of the most frequently used words in each category, clearly highlighting contrasts. The use of distinct primary colors ensures clear differentiation and likely draws the viewer's attention effectively. The chart is also clearly labeled, with simple axis titles and an interactive (toggleable) legend that ensures the viewer quickly understands what the visualization represents, and also has the choice to view each graph separately as well. The user can hover over each bar to view the word and its corresponding frequency difference count when needed.

However, the visualization might benefit from additional context (like what I provided above), such as a brief explanation of how these words were selected or what the implications of these differences are. Although this is actually included in the dashboard version, I did not want to overburden the visualization itself with text and keep it as clean as possible. I could also have visualized it a different way—instead of using the frequency difference, I could simply use the raw frequency of AI and Human word usage and show that in a stacked bar chart to visually show the difference rather than visualizing the difference calculation. There is a chance that would be more intuitive to look at, but as of now, I am satisfied with this visualization for the scope of this project.

# Critical Analysis

## Challenges and Limitations

I underestimated how challenging it would be to derive meaningful conclusions from a trigram analysis without utilizing more complex NLP methods. A problem I ran into consistently when I was working through my original hypothesis was that the engrams and trigrams kept showing me phrase patterns relating to the *topic* of the essays (e.g. “limit car usage”) rather than the typical AI phrasing (e.g. In conclusion, the evidence points to…) I was looking to compare with less formulaic human phrasing. Therefore, the visualizations also did not tell any kind of meaningful story. Since this dataset was tailored for LLM usage, it may not have been the ideal choice—though its subtitle, “Detect AI Generated Text” and its inclusion of a number of AI and human-written essays on a variety of subjects made me think otherwise. Since the analysis kept focusing on the topics of the essays, I decided to make that my focus, which I think did make the entire project more meaningful and tell a better story (that was in the scope of this specific project and time constraints).

Another challenge I ran into after shifting my focus was the wrongful categorization of topics when using solely LDA for topic modelling. When I created the initial word clouds for each topic, I saw certain words and engrams in the wrong category (e.g. electoral college in the Education category instead of Politics). I took the help of AI models for suggestions on ways to minimize this error, and the one that seemed the best to me was specifying certain keywords for more accurate topic classification. This worked well, as it utilized the other NLP methods I conducted in tandem with this manual method, and the word clouds were much more accurate, meaning the other visualizations would also use more accurate data since my visualizations were now all topic-focused.

One major struggle was the dashboard. I wanted to make it much more interactive and add more user controls, but despite multiple different attempts, including using the lab code from the text analysis dash, manual coding, looking up documentation, and taking help from AI, I had no success. A few features I was trying to include were:

* A slider for adjusting word cloud frequency threshold to allow users to filter out less meaningful words.
* A checkbox to show/hide certain topics in the bar chart & treemap so users can choose which topics they want to compare.
* A hover-over tooltip in the AI vs. Human Word Frequency Difference Chart that shows exact AI and Human word frequencies and also % difference instead of just absolute counts

First, the features were actually showing up in the dashboard but were not functional, and after trial and error, the data in the graphs were not showing up altogether. I decided to then just focus on making the dashboard look slightly more visually appealing and less basic by changing its typography and adding descriptions for the graphs.