Exploring Dataset Report

Full Name: Nguyễn Đình Thiên Lộc

Student ID: 24125093

Course: Scientific Method (SC203)

Date: 24/10/2025

1. Topic Selection

Selected Topic: Evaluating AI Detection Accuracy through Nominalization Frequency in Human vs AI-Generated Essays

Input: A dataset of essays labeled as human-written (0) or AI-generated (1) obtained from Kaggle.

Output: Quantitative comparison of word count and noun (nominalization) frequency between human and AI-generated essays.

Scope: Included: Linguistic exploration based on noun and word counts. Excluded: Other stylistic or syntactic features such as sentence complexity, punctuation, or semantic coherence.

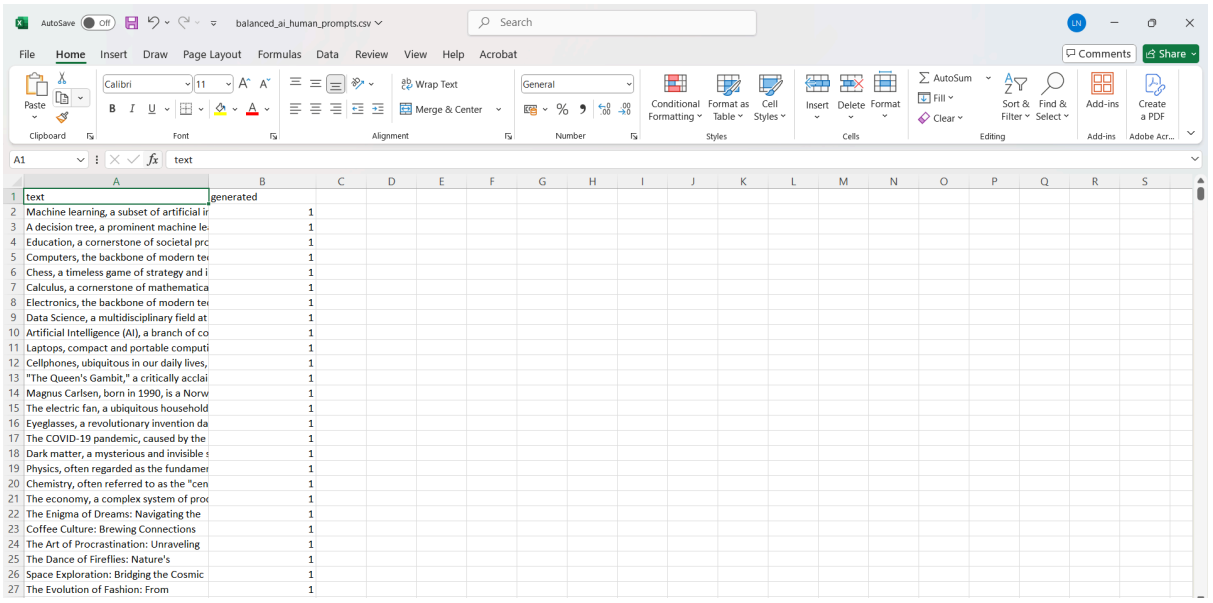
2. Dataset Description

Dataset Title: Human vs AI Generated Essays

Source: https://www.kaggle.com/datasets/navjotkaushal/human-vs-ai-generated-essays?resource=download

Overview: The dataset contains short essays written by both humans and AI systems. Each record includes a text column (the essay) and a label column (0 = Human, 1 = AI). The dataset is balanced, allowing fair comparison between human and AI-written texts.

FIGURE 1: ORIGINAL DATASET



	A	B
1	text	generated
2	Machine learning, a subset of artificial ir	1
3	A decision tree, a prominent machine le	1
4	Education, a cornerstone of societal pro	1
5	Computers, the backbone of modern te	1
6	Chess, a timeless game of strategy and i	1
7	Calculus, a cornerstone of mathematica	1
8	Electronics, the backbone of modern te	1
9	Data Science, a multidisciplinary field at	1
10	Artificial Intelligence (AI), a branch of co	1
11	Laptops, compact and portable computi	1
12	Cellphones, ubiquitous in our daily lives,	1
13	"The Queen's Gambit," a critically acclai	1
14	Magnus Carlsen, born in 1990, is a Norw	1
15	The electric fan, a ubiquitous household	1
16	Eyeglasses, a revolutionary invention da	1
17	The COVID-19 pandemic, caused by the	1
18	Dark matter, a mysterious and invisible s	1
19	Physics, often regarded as the fundamer	1
20	Chemistry, often referred to as the "cen	1
21	The economy, a complex system of proc	1
22	The Enigma of Dreams: Navigating the	1
23	Coffee Culture: Brewing Connections	1
24	The Art of Procrastination: Unraveling	1
25	The Dance of Fireflies: Nature's	1
26	Space Exploration: Bridging the Cosmic	1
27	The Evolution of Fashion: From	1

3. Data Exploration and Understanding

I used ChatGPT to generate Python code to automatically detect, load, and preview the dataset. The script displays the first few rows, column names, and category counts, proving that the dataset was actively explored.

4. Data Manipulation and Analysis

To manipulate the dataset, I added two features: word_count (total words per essay) and noun_count (total nouns per essay using NLTK POS tagging). This allows quantitative analysis of nominalization differences between human and AI essays.

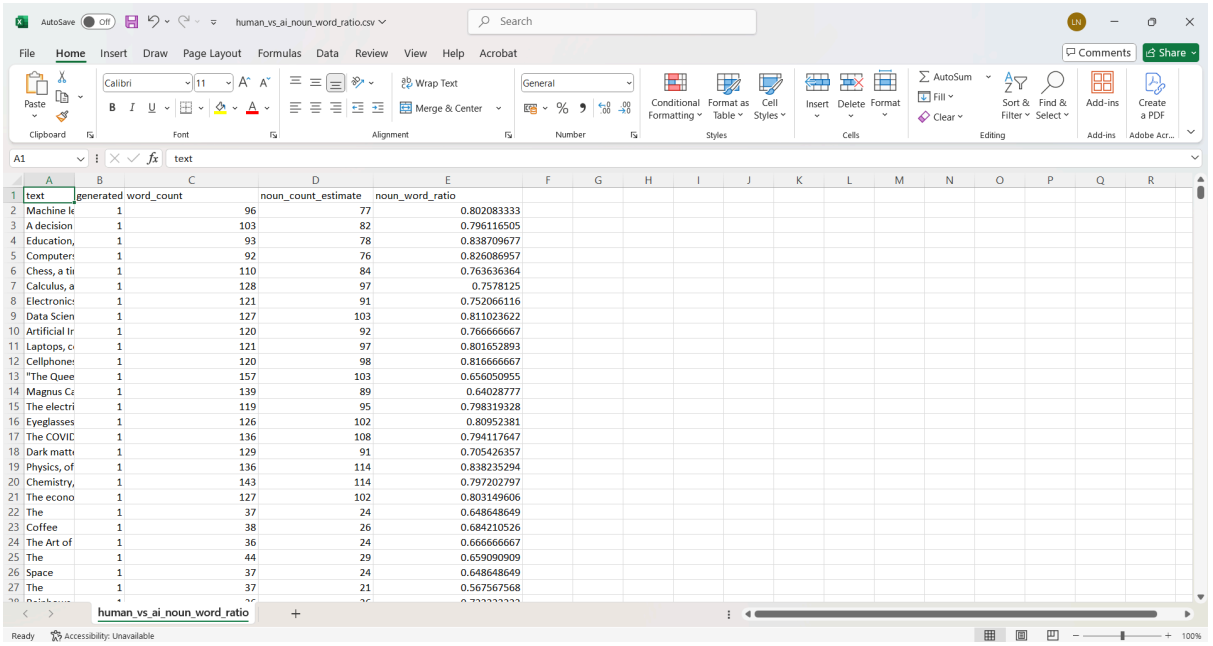
Python implementation:

```
# Load and preview dataset

df = pd.read_csv("human_vs_ai_essays.csv")

print(df.head())

# Add linguistic features

df['word_count'] = df['text'].apply(lambda x: len(x.split()))

df['noun_count'] = df['text'].apply(count_nouns)
```

```
# Calculate ratio and plot results

df['noun_word_ratio'] = df['noun_count'] / df['word_count']
```

The full script (≈200 lines) automatically detects the dataset, standardizes column names, computes noun and word counts, and generates bar charts for average noun and word frequencies.

FIGURE 2: DATASET AFTER MANIPULATION



text	generated	word_count	noun_count_estimate	noun_word_ratio
Machine le	1	96	77	0.802083333
A decision	1	103	82	0.796116505
Education,	1	93	78	0.838709677
Computers	1	92	76	0.826086957
Chess, a ti	1	110	84	0.763636364
Calculus, a	1	128	97	0.7578125
Electronics	1	121	91	0.752066116
Data Scien	1	127	103	0.811023622
Artificial Ir	1	120	92	0.766666667
Laptops, c	1	121	97	0.801652893
Cellphones	1	120	98	0.816666667
"The Quee	1	157	103	0.656050955
Magnus Ca	1	139	89	0.64028777
The electri	1	119	95	0.798319328
Eyeglasses	1	126	102	0.80952381
The COVID	1	136	108	0.794117647
Dark matte	1	129	91	0.705426357
Physics, of	1	136	114	0.838235294
Chemistry,	1	143	114	0.797202797
The econo	1	127	102	0.803149606
The	1	37	24	0.648648649
Coffee	1	38	26	0.684210526
The Art of	1	36	24	0.666666667
The	1	44	29	0.659090909
Space	1	37	24	0.648648649
The	1	37	21	0.567567568

5. Findings and Preliminary Observations

After processing the dataset and computing both the **word count** and **noun count**, I generated basic descriptive figures to explore potential trends between human-written and AI-generated essays.

An example preview of the dataset after manipulation is shown in **Figure 2**, which confirms that additional linguistic columns were successfully generated.

6. Next Steps and Considerations

This stage of the project focused on **dataset exploration and preparation**. The goal was to examine the available columns, calculate basic linguistic statistics, and visualize preliminary metrics like word and noun counts.

Future steps will include:

1. **Expanding analysis** to include the *noun-to-word ratio* (noun density) as a normalized metric.

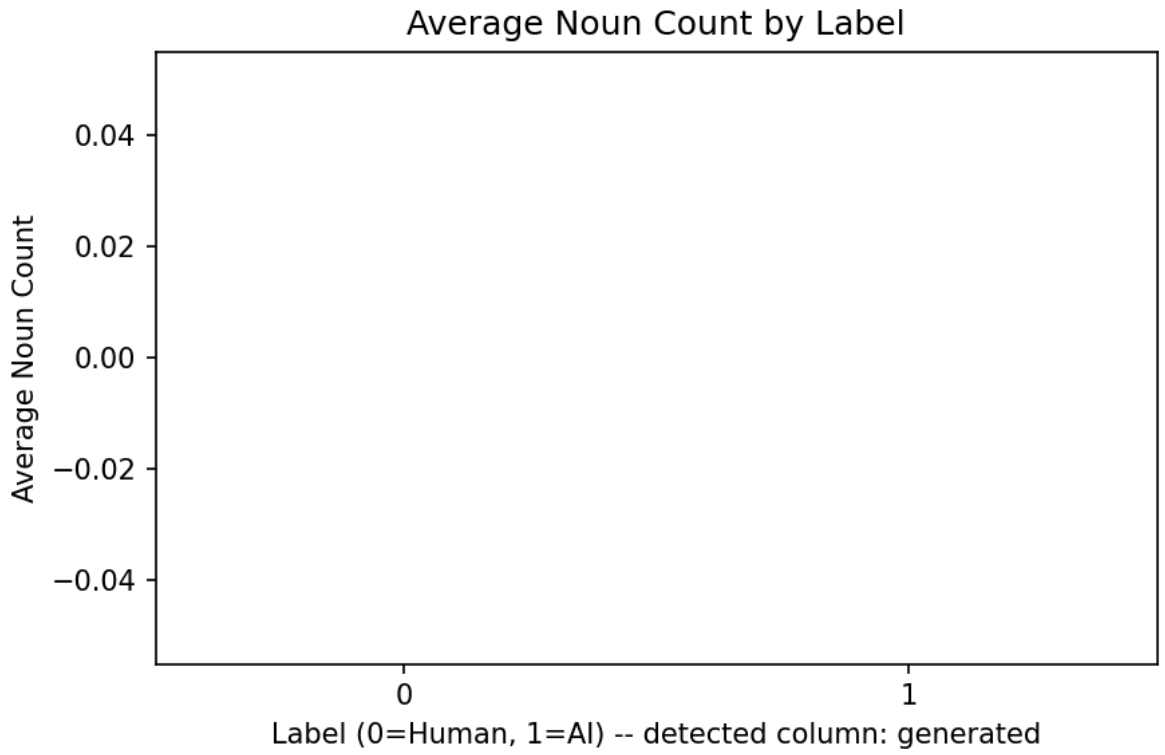
2. **Comparing categories statistically** using averages and standard deviations to identify meaningful differences.

3. **Testing assumptions** about whether nominalization frequency or noun density correlate with AI-generated writing styles.

At this point, no conclusions are drawn about the relationship between these variables. The figures and results serve only as **early exploratory findings** that will guide more detailed analysis in subsequent stages.

I have managed to generate some bar charts but they are still faulty due to logic bugs in coding in Python.

FIGURE 3: A POTENTIAL CHART FORMAT FOR ANALYZING IN THE FUTURE



7. References

Kaushal, N. (2023). Human vs AI Generated Essays Dataset. Kaggle. https://www.kaggle.com/datasets/navjotkaushal/human-vs-ai-generated-essays?resource=download

Bird, S., Loper, E., & Klein, E. (2009). Natural Language Processing with Python. O'Reilly Media.