Generative AI Misinformation Research Final Report

STAT/DSDA 1010; Data Science and Society Using R

Karissa Wong

1 Introduction

1.1 Overview of Data

This research project primarily refers to its corresponding Shiny Dashboard and uses the Gen AI Misinformation Detection Data (2024-2025) dataset provided by Atharva Soundankar on Kaggle. The dataset contains realistic simulations of news articles and social media posts (referred to as a "post" in this project) that circulated 2024-2025 and were labeled as potential AI-generated misinformation. Each entry represents a post and its corresponding attributes, which can be categorized into time, location, author, attributes, credibility, and misinformation.

1.1.1 Target Variables

For the purpose of this project, the variables that will be focused on are **country, city, token count, readability score, sentiment score, toxicity score, model signature, engagement,** and **if the post is misinformation or not.** A brief explanation of each variables can be found below:

- **Country:** The country where the post originated.
- **City:** The city, corresponding with its respective country, where the post originated.
- **Token count:** In AI, a token is a unit of data that comes from breaking down larger words or phrases. AI models use tokens to process responses and have limits to how many tokens can be processed at a time. Thus, more tokens result in outputs that take more time.
- **Readability score:** A numerical value measuring how easy it is to understand a piece of text, scoring from 0 to 100. A score closer to 100 indicates that the text is very easy to understand.
- **Sentiment score:** A numerical value representing the overall emotion expressed in a piece of text, scoring from -1 to 1. A score closer to -1 indicates an overall negative tone, closer to 0 indicates a neutral tone, and closer to 1 indicates an overall positive tone.
- **Toxicity score:** A numerical value measuring how harmful or offensive the content of the text is. A score less than 0.5 indicates toxicity and a score greater than 0.5 indicates nontoxicity.
- **Model Signature:** The type of writing style a post most closely reflects. The three model signatures identified were GPT-like (AI), human, and unknown.
- **Engagement:** A numerical value representing the sum of likes, shares, comments, and views a post has received.
- **Misinformation:** Indicates whether the content in the post is misinformation or authentic information.

1.2 Purpose

Some common attributes distinguish AI writing as its presence has increased over the years. AI-generated text often has **higher readability scores** than human writing because algorithmic systems favor more **simplistic writing**, resulting in **lower token counts** to produce text faster. AI-generated text also **lacks emotional depth** has due to its computational structure, resulting in more **neutral sentiment and toxicity scores** (0 and 0.5, respectively).

The goal of this research project is to **study the impact of the increase in Gen AI on misinformation in newswriting and social media posts and to test common attributes of AI writing.** The **primary focus** of this project is to answer the questions of what factors cause text to be marked as GPT-like writing and the probability that text marked as GPT-like is actually misinformation. Additionally, this project focuses on which countries produce the most AI-generated information and how much engagement each post receives based on their model signature. This analysis can help explore potential patterns in content being marked as AI misinformation, and if there is a need for change in current media writing structure to avoid this problem.

2 Analysis

2.1 Detected Misinformation Around the World

In the **"Detected Misinformation Around the World"** tab, the map shows the **unique locations** each post originated from with a filter for model signature, country, and city. The map corresponds to a frequency bar chart, showing the number of misinformation and authentic information posts by model signature, country, city, or all three factors.

2.1.1 Findings

Figure 1: AI Posts by Country

Based on this sample of data, there is **not a significant difference in the number of AI posts produced by each country.** However, it is worth noting that every post originated from metropolitan areas, aligning with the trend of larger cities adopting AI because of its ability to provide more funding and research.

2.2 Engagement with Content

In the **"Engagement with Content"** tab, the column chart shows the **relationship between engagement and posts with misinformation versus posts with authentic information.** The chart also has filters for model signature, country, and city to compare engagement with specific model signatures and engagement across countries.

2.2.1 Findings

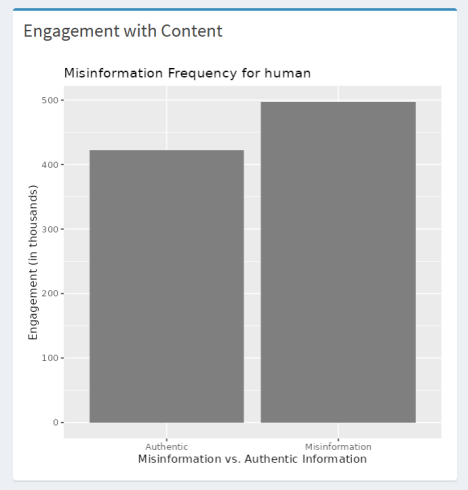

Figure 2: Engagement with Human Content

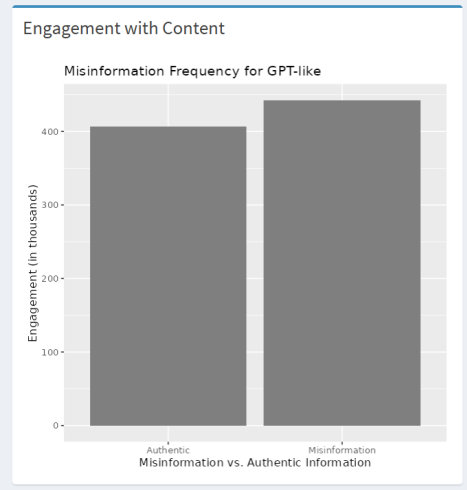

Figure 3: Engagement with AI Content

The column chart shows that people overall **engage with more human content** (919,439 cases of engagement) than AI content (849,052 cases of engagement), showing the continued dominance of human content in social media.

2.3 Misinformation Detection Frequency

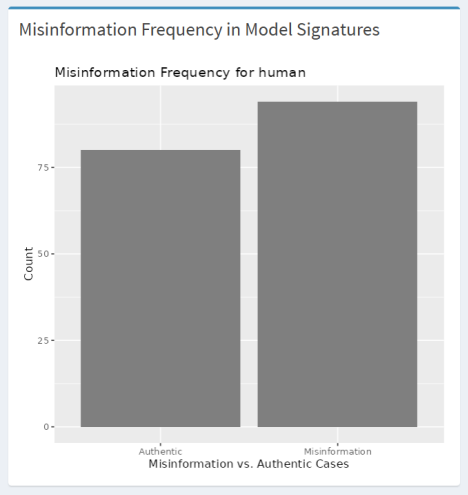

Figure 4: Misinformation in Human Posts

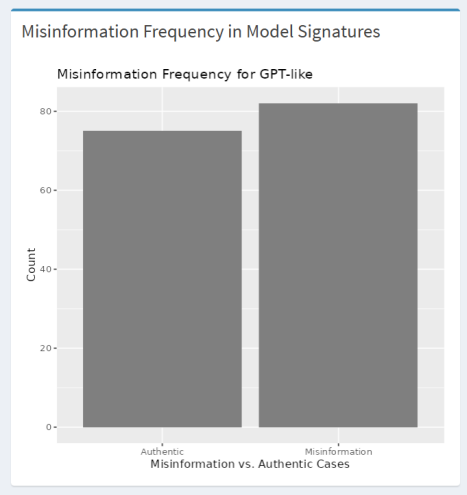

Figure 5: Misinformation in AI Posts

In the **"Misinformation Detection Frequency"** tab, the frequency bar chart **compares the count of misinformation content and authentic content by model signature.** In a sample of 157 GPT-like posts and 174 human posts, the amount of **misinformation posts were slightly higher in human posts** ($p_1 = 54\%$) compared to GPT-like posts ($p_2 = 52.2\%$).

2.3.1 Testing for Statistical Significance

A **two-proportion Z-test** was conducted to see if there is a significant difference between p_1 and p_2. he **null hypothesis** was there is no difference between the amount of misinformation posts in each model signature and the **alternative hypothesis** was the proportion of misinformation in GPT-like posts is greater than in human posts.

```
    2-sample test for equality of proportions with continuity correction

data:  c(75, 80) out of c(157, 174)
X-squared = 0.04677, df = 1, p-value = 0.4144
alternative hypothesis: greater
95 percent confidence interval:
 -0.07846371  1.00000000
sample estimates:
   prop 1    prop 2
0.4777070 0.4597701
```

The test produced a **p-value of around 0.4144**, meaning that there is not enough evidence to conclude a statistically significant difference given a significance level of 0.05.

2.4 Attributes of Misinformation Detection

Figure 6: Post Attributes

In the **"Attributes of Misinformation Detection"** tab, each boxplot shows the **summary distribution of each attribute** in posts with GPT-like and human model signatures. The summary statistics show the quartiles, median, mean, minimum, maximum, and spread of each attribute. This helps summarize the distribution of **token counts, toxicity score, sentiment score,** and **readability score** in GPT-like and human posts. A **two-sample T-test** was conducted for each attribute to see if there is a significant difference between the means of the GPT-like and human post group.

2.4.1 Readability Score Findings

For the readability score, the **null hypothesis** was there is no difference between the means of the GPT-like and human group and the **alternative hypothesis** was there is a difference between the means. The mean and standard deviation for the **GPT-like** group was 54.96 and 14.95, respectively. The mean and standard deviation for the **human** group was 55.29 and 14.74, respectively.

```
    Welch Two Sample t-test

data:  ai_readability and human_readability
t = -0.19844, df = 324.55, p-value = 0.8428
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.540118  2.891387
sample estimates:
mean of x mean of y
 54.96414  55.28851
```

The test produced a **p-value of around 0.8428**, meaning that there is not enough evidence to conclude there is a statistically significant difference given a significance level of 0.05.

2.4.2 Sentiment Score Findings

For the sentiment score, the **null hypothesis** was there is no difference between the means of the GPT-like and human group and the **alternative hypothesis** was there is a difference between the means. The mean and standard deviation for the **GPT-like** group was -0.02 and 0.57, respectively. The mean and standard deviation for the **human** group was 0.02 and 0.59, respectively.

```
    Welch Two Sample t-test

data:  ai_sentiment and human_sentiment
t = -0.70275, df = 327.4, p-value = 0.4827
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.16977714  0.08040486
sample estimates:
  mean of x   mean of y
-0.02137580  0.02331034
```

The conclusions of this test was a **p-value of around 0.4827**, meaning that there is not enough evidence to conclude there is a statistically significant difference given a significance level of 0.05.

2.4.3 Toxicity Score Findings

For the toxicity score, the **null hypothesis** was there is no difference between the means of the GPT-like and human group and the **alternative hypothesis** was there is a difference between the means. The mean and standard deviation for the **GPT-like** group was 0.51 and 0.31, respectively. The mean and standard deviation for the **human** group was 0.50 and 0.28, respectively.

```
    Welch Two Sample t-test

data:  ai_toxic and human_toxic
t = 0.32419, df = 319.02, p-value = 0.746
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05358029  0.07472137
sample estimates:
mean of x mean of y
0.5091338 0.4985632
```

The conclusions of this test was a **p-value of around 0.6204**, meaning that there is not enough evidence to conclude there is a statistically significant difference given a significance level of 0.05.

2.4.4 Token Count Findings

For the token count, the **null hypothesis** was there is no difference between the means of the GPT-like and human group and the **alternative hypothesis** was there is a difference in means. The mean and standard deviation for the **GPT-like** group was 34.22 and 20.43, respectively. The mean and standard deviation for the **human** group was 35.84 and 19.60, respectively.

```
    Welch Two Sample t-test

data:  ai_token and human_token
t = -0.73806, df = 322.3, p-value = 0.461
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.968512  2.711978
sample estimates:
mean of x mean of y
 34.21656  35.84483
```

The conclusions of this test was a **p-value of around 0.461**, meaning that there is not enough evidence to conclude there is a statistically significant difference given a significance level of 0.05.

3 Conclusion

Based on the analysis of this dataset, there is **no significant difference** between AI and human writing and the amount of misinformation each model produces. There is not enough evidence to show that the **average difference** between the attributes of AI and human posts is significantly different, and there is **no definitive factor** that causes text to be labeled as GPT-like. Given these results, a conclusion cannot be confidently made on whether there is a need to change current media writing structure in the age of AI.

The conclusion of this research project emphasizes how **narrow the gap between AI and human writing has become** and how crucial it is to understand **how to identify AI content**, as well as **how to identify misinformation**. AI content generally has a formulaic structure, repeated phrases, and a more consistent tone compared to human content. It is also crucial to assess the credibility of sources, as misinformation can appear in any type of content. AI is prone to citing outdated or incorrect sources when providing information, but it is still important to verify information regardless of whether it comes from AI or human authors.