STA202 - Final Report - Group 14

Survey and Analysis of Ability to Identify AI Generated Texts

Group Member Name	Enrollment Number
Prasham Pinkesh Shah	AU2120137
Gautam Nipesh Shah	AU2120227
Shubham Shah	AU2120064
Mujtaba Jafri	AU2120119

Abstract

This study used statistical analysis to explore people's ability to distinguish between text created by artificial intelligence (AI) and text written by humans. A group of 158 participants of different education levels, ages, and sex evaluated eight questions. The findings indicate that on average, participants were generally good at identifying the AI-generated text. Those with undergraduate-level education or less were able to correctly identify the text 73.9% of the time, while those with higher education did so 64.1% of the time. Participants under 24 years old had a 70% correct identification rate, compared to 76.92% for those over 24. Females had a higher correct identification rate (74.6%) than males (69.47%). These results highlight demographic differences in the ability to detect AI-generated content and suggest a need for further research into the underlying factors.

Problem

We examine the growing issue of AI-generated text that closely resembles human writing. This raises concerns about the spread of misinformation and trust in content, as distinguishing AI-generated text from human-authored material can be challenging. The study investigates people's capacity to differentiate between AI-generated and human-written text, considering the potential implications for societal trust and information reliability.

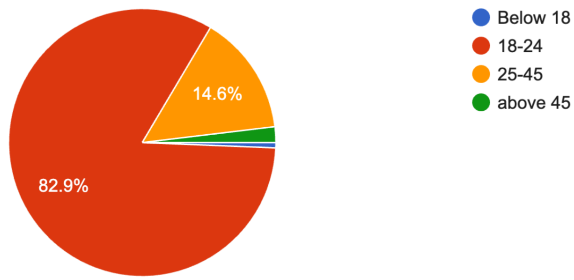
Plan

Our plan was to conduct a survey and perform analysis on the data collected with data points belonging to varied groups across categories. We look at age, education level and sex.
The survey includes eight questions, with half generated by artificial intelligence (AI) and the other half written by Humans . The question categories were chosen to reflect the Indian context, encompassing religion, political campaigning, scientific facts, and female hygiene.

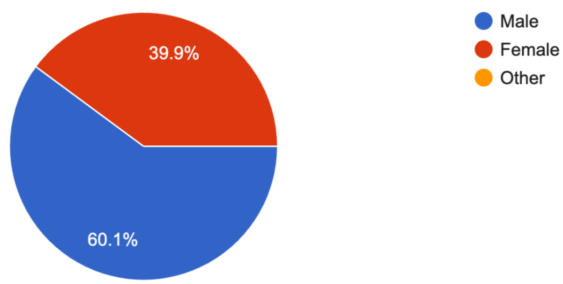
Data

This project utilizes a primary dataset collected through a survey. A total of 158 respondents participated in the survey, with a majority being college students familiar with AI-generated text. All the categorical variables were later converted into binary to reduce skewness, these were UG and Above UG for Education, below 24 and above 24 for Age.

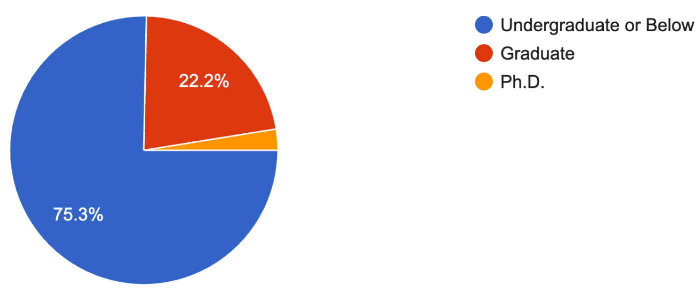
Age
158 responses



Sex
158 responses



Educational Background
158 responses



Exploratory Data Analysis

Dataset summary

	Human Score	AI Score	Total Score	Sex_Encoded	High_Ability
count	158.000000	158.000000	158.000000	158.000000	1258.000000
mean	2.177215	2.139241	4.316456	1.398734	0.054054
std	0.974313	0.986975	1.497721	0.491195	0.226214
min	0.000000	0.000000	0.000000	1.000000	0.000000
25%	2.000000	1.250000	3.000000	1.000000	0.000000
50%	2.000000	2.000000	4.000000	1.000000	0.000000
75%	3.000000	3.000000	5.000000	2.000000	0.000000
max	4.000000	4.000000	8.000000	2.000000	1.000000

Central Tendencies

Median Total Score	4
Mean Total Score	4.316455696
Mode Total Score	4

Education vs Ability (High Ability : >=4 correct)

	Low Ability	High Ability	Total	High %
UG	31	88	119	73.9495798 3
Above UG	14	25	39	64.1025641

Age vs Ability (High Ability : >=4 correct)

	Low Ability	High Ability	Total	High %
Under 24	39	93	132	70.4545454 5
Above 24	6	20	26	76.9230769 2

Gender vs Ability (High Ability : >=4 correct)

	Low Ability	High Ability	Total	High %
Male	29	66	95	69.4736842 1
Female	16	47	63	74.6031746

Females were generally better at identifying Human text

Human				
	Male correct	**Female correct**	**Ratio Male**	**Ratio Female**
3rd	51	35	0.5368421053	**0.5555555556**
4th	50	38	0.5263157895	**0.6031746032**
5th	49	33	0.5157894737	**0.5238095238**
8th	50	38	0.5263157895	**0.6031746032**

7th question was related to menstrual hygiene (77% of all the females who attempted got it right)

AI				
	Male correct	**Female correct**	**Ratio male**	**Ratio Female**
1st question	57	33	**0.6**	0.5238095238
2nd question	45	34	0.4736842105	**0.5396825397**
6th question	59	27	**0.6210526316**	0.4285714286
7th question	34	49	0.3578947368	**0.7777777778**

Code

analysis.R

```
data = read.csv('Form_Responses.csv', sep=",", header = TRUE)
# converting categorical string variables to factors
data$Sex = as.factor(data$Sex)
data$AGE..below.and.above.24. = as.factor(data$AGE..below.and.above.24.)
data$Education.above.Ug..below.UG. = as.factor(data$Education.above.Ug..below.UG.)

# Drop the 'Age' and 'Sex_Encoded' columns if it exist
data = data[, !(names(data) %in% c("Timestamp","Email.Address","Age", "Sex_Encoded"))]

# Create a High Ability columns
data$High_Ability = ifelse(data$Total.Score >= 4, "Yes", "No")

#converting it to a factor
data$High_Ability = as.factor(data$High_Ability)
# Check the structure of the data frame after adding the new column
str(data)

#performing logistic regression
logistic_ALL = glm(High_Ability ~ Sex+ AGE..below.and.above.24.+ Education.above.Ug..below.UG. , data = data, family=binomial)
summary(logistic_ALL)

logistic_sa = glm(High_Ability ~ Sex+ AGE..below.and.above.24., data = data, family=binomial)
summary(logistic_sa)

# predict the high ability based on the model
data$predicted_high_ability = predict(logistic_ALL, type = "response") > 0.5

# Creating the confusion matrix using table function
confusion_matrix = table(data$High_Ability, data$predicted_high_ability)

# Print the confusion matrix
print(confusion_matrix)

# Extract values from the confusion matrix
true_positives = confusion_matrix["Yes", "TRUE"]
false_positives = confusion_matrix["No", "TRUE"]
false_negatives = confusion_matrix["Yes", "FALSE"]
true_negatives = confusion_matrix["No", "FALSE"]

# calculate the metrics
precision = true_positives / (true_positives + false_positives)
recall = true_positives / (true_positives + false_negatives)
accuracy = (true_positives + true_negatives) / sum(confusion_matrix)
f1_score = 2 * (precision * recall) / (precision + recall)

# Print the results
cat("Precision:", precision)
cat("Recall:", recall)
cat("Accuracy:", accuracy)
cat("F1-score:", f1_score)

## now we can plot the data
predicted.data = data.frame(
  probability.of.High_Ability =logistic_ALL$fitted.values,
  High_Ability=data$High_Ability)

# sort the dataframe from low to high probabilities
predicted.data = predicted.data[
  order(predicted.data$probability.of.High_Ability, decreasing=FALSE),]

# add new column for rank
predicted.data$rank = 1:nrow(predicted.data)

# load libraries
library(ggplot2)

# using geom_point plot the data
ggplot(data=predicted.data, aes(x=rank, y=probability.of.High_Ability)) +
  geom_point(aes(color=High_Ability), alpha=1, shape=4, stroke=2) +
  xlab("Index") +
  ylab("Predicted probability of High Ability")

ggsave("plot.pdf")
```

Model Summary:

```
Call:
glm(formula = High_Ability ~ Sex + AGE..below.and.above.24. +
    Education.above.Ug..below.UG., family = binomial, data = data)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                        1.2404     0.5030   2.466   0.0137 *
SexMale                           -0.2161     0.3744  -0.577   0.5637
AGE..below.and.above.24.Under 24  -1.2779     0.6814  -1.875   0.0607 .
Education.above.Ug..below.UG.UG    1.2030     0.5469   2.200   0.0278 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 188.79  on 157  degrees of freedom
Residual deviance: 183.06  on 154  degrees of freedom
AIC: 191.06

Number of Fisher Scoring iterations: 4
```

Model and Results

Model Used: Logistic Regression

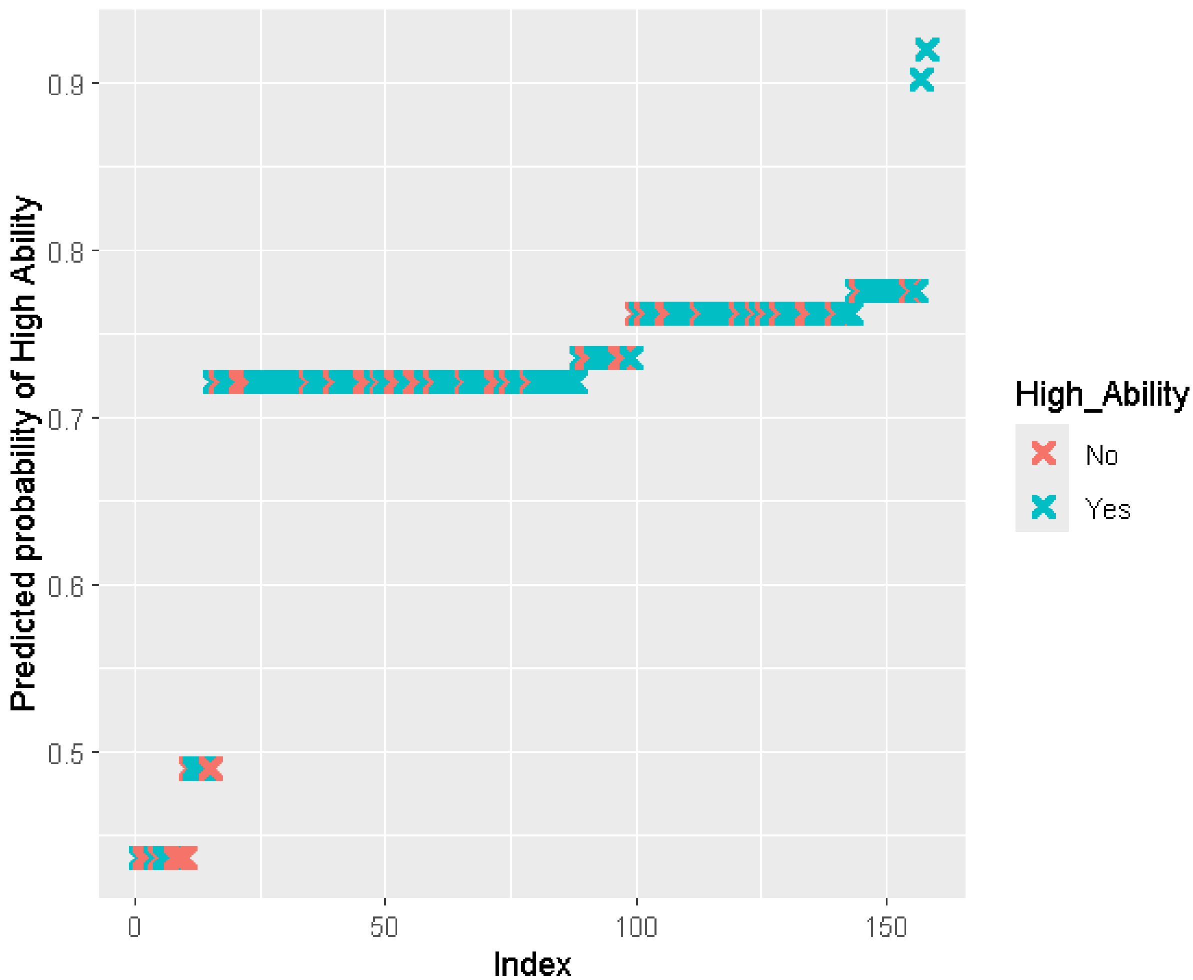
Logistic regression is a well-suited choice for several reasons:

- **Binary Dependent Variable:** Our dependent variable, "High_Ability" (Yes/No), is categorical with only two possible outcomes. Logistic regression is specifically designed to model the relationship between a binary outcome and one or more independent variables.
- **Predict Probabilities:** Unlike linear regression, logistic regression allows us to estimate the probability of an observation belonging to a specific category (Yes, in this case) based on the independent variables.
- **Interpretation of Coefficients:** The coefficients obtained from the logistic regression model represent the log-odds change in the dependent variable for a one-unit increase in the independent variable (holding other variables constant). By interpreting these coefficients and their significance levels, we can gain insights into which factors are statistically significant predictors of high AI text identification ability and the direction of their influence.

Results

High Ability: Total Score >=4

Prediction Graph:



Confusion Matrix:

```
> print(confusion_matrix)

      FALSE TRUE
No        8   37
Yes       7  106
```

Evaluation Metrics:

```
> # Print the results
> cat("Precision:", precision)
Precision: 0.7412587
> cat("Recall:", recall)
Recall: 0.9380531
> cat("Accuracy:", accuracy)
Accuracy: 0.721519
> cat("F1-score:", f1_score)
F1-score: 0.828125
```

Confidence Interval of coefficient Estimation:

```
> confint(logistic_ALL)
Waiting for profiling to be done...
                                        2.5 %
(Intercept)                         0.3109561
SexMale                            -0.9647614
AGE..below.and.above.24.Under 24 -2.6793485
Education.above.Ug..below.UG.UG     0.1369962
                                       97.5 %
(Intercept)                        2.31395223
SexMale                            0.51019664
AGE..below.and.above.24.Under 24 0.01658124
Education.above.Ug..below.UG.UG    2.31159716
```

```
> confint.default(logistic_ALL)
                                        2.5 %
(Intercept)                         0.2545792
SexMale                            -0.9498456
AGE..below.and.above.24.Under 24 -2.6133654
Education.above.Ug..below.UG.UG     0.1311060
                                       97.5 %
(Intercept)                        2.22628414
SexMale                            0.51760942
AGE..below.and.above.24.Under 24 0.05766226
Education.above.Ug..below.UG.UG    2.27494210
```

Odds Ratio:

```
> exp(coef(logistic_ALL))
                        (Intercept)
                          3.4571055
                            SexMale
                          0.8056402
AGE..below.and.above.24.Under 24
                          0.2786353
 Education.above.Ug..below.UG.UG
                          3.3301723
```

Interpretation

- Mens odds of being able to correctly identify AI and Human generated text is smaller than that of women by a factor of 0.805.
- People under 24 years old have a 0.28 times lower odds of having high ability compared to people over 24 years old.
- People with a UG or below education level have 3.33 times higher odds of having high ability compared to people with an education level above UG.

Conclusion

In conclusion, our study underscores the significant challenge many individuals face in distinguishing between AI-generated text and human-authored content. Our analysis of 158 participants representing diverse demographics, including education level, age, and sex, revealed only moderate accuracy in identifying AI-generated text on average. Notably, participants with lower education levels tended to perform slightly better than those with higher education levels, and younger individuals showed slightly better performance than older participants. Additionally, females outperformed males in correctly identifying AI-generated content. These findings highlight the importance of understanding demographic differences in text detection abilities and the potential implications for societal trust and information reliability. Moving forward, further research into the underlying factors influencing individuals' capacity to differentiate between AI-generated and human-written text is essential. Such research is crucial for developing strategies to mitigate the potential negative consequences of AI-generated content on trust in information and the reliability of online discourse. Our study contributes valuable insights to the broader discourse on the ethical and societal implications of AI technology, particularly in the context of content creation and dissemination. It provides valuable guidance for policymakers, educators, and technologists aiming to address the evolving landscape of digital communication and information dissemination.

References

1. How To Run Logistic Regression In R. How to run logistic regression in R. (2024). https://www.nbshare.io/notebook/756634696/How-To-Run-Logistic-Regression-In-R/
2. Chugh, V. (2023, March 17). *Logistic regression in R tutorial*. DataCamp. https://www.datacamp.com/tutorial/logistic-regression-R
3. Bobbitt, Z. (2023, March 11). *How to perform logistic regression in Excel*. Statology. https://www.statology.org/logistic-regression-excel/