AI or Human? A Machine Learning Approach to Text Classification for Statistical Courses

By Andrew John J, Roshni Bhowmik, Mahnoor Sheikh, Ab Basit Syed Rafi

Instructor – Savvy Barnes *Course: STT 811*

Executive Summary

This project explores the classification of human-generated and AI-generated text in a statistical education context. Using a dataset of responses to 116 statistics questions, we conducted extensive feature engineering, exploratory data analysis, and built multiple machine learning models, including BERT, to distinguish between text origins. The best-performing model achieved an accuracy of 85%.

Introduction

With the rise of large language models like ChatGPT, ensuring academic integrity has become increasingly critical. This project develops a predictive system to classify text as either human- or AI-generated using traditional machine learning models and deep learning.

Dataset Overview

The dataset consists of 2,239 entries covering:

- Question: The original question.
- Human Response: Answer written by a human.
- AI Response: Answer generated by an LLM.

Post-cleaning, 1,993 usable rows remained.

Preprocessing & Feature Engineering

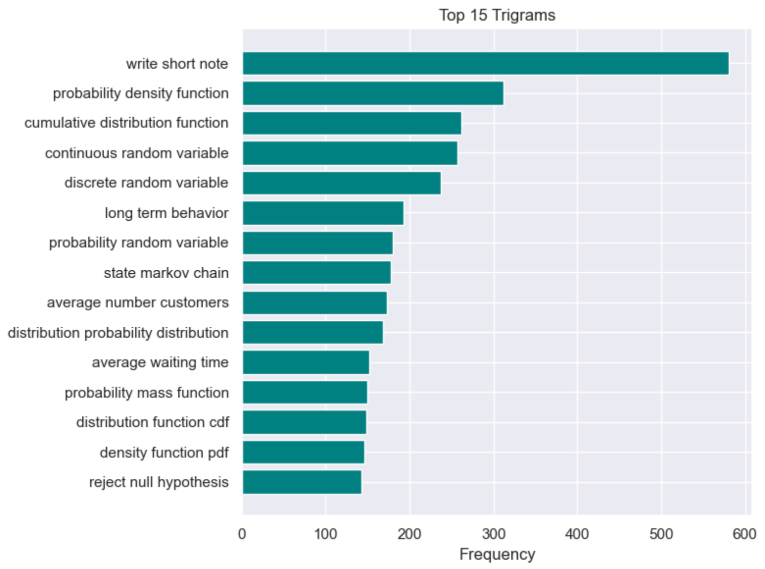
Key steps included:

- Dropping rows with missing entries.
- Creating numerical features (text length, special character counts). Normalizing them using standardscaler().
- Text cleaning (lowercasing, removing punctuation, tokenization). Removal of stop words.
- Creating a sparse document-term matrix using CountVectorizer.
- Dimensionality reduction with PCA (95% variance captured by 482 components).

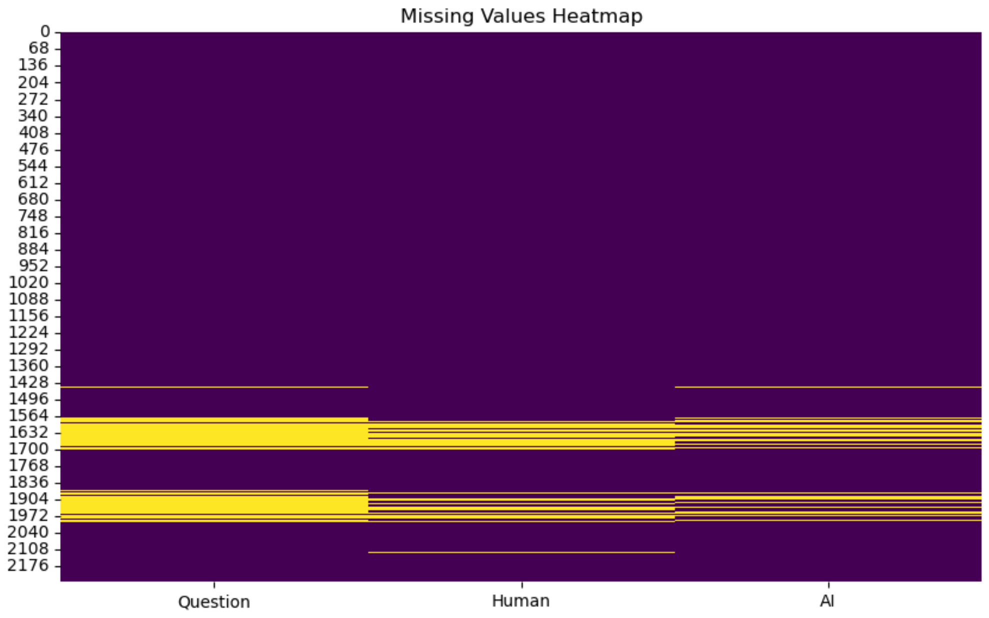
Exploratory Data Analysis (EDA)

Highlights:

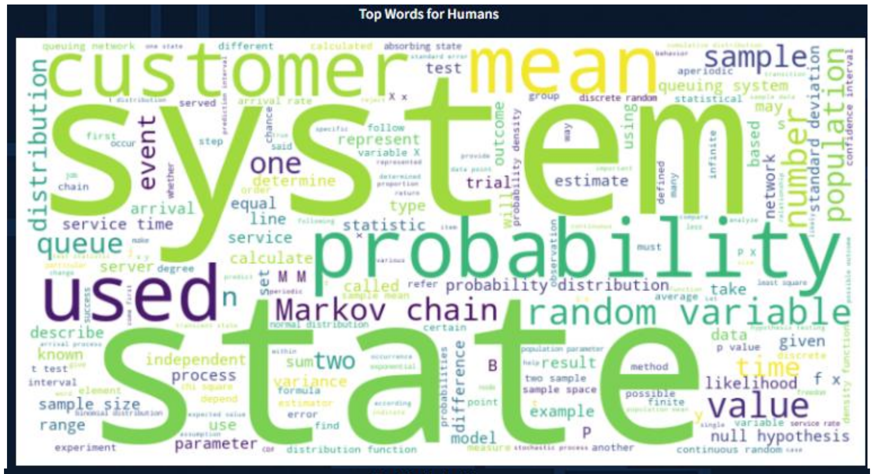
- Top Trigrams: The bar chart displays the top 15 most frequent trigrams (three-word phrases) in a text corpus, with "write short note" being the most common.

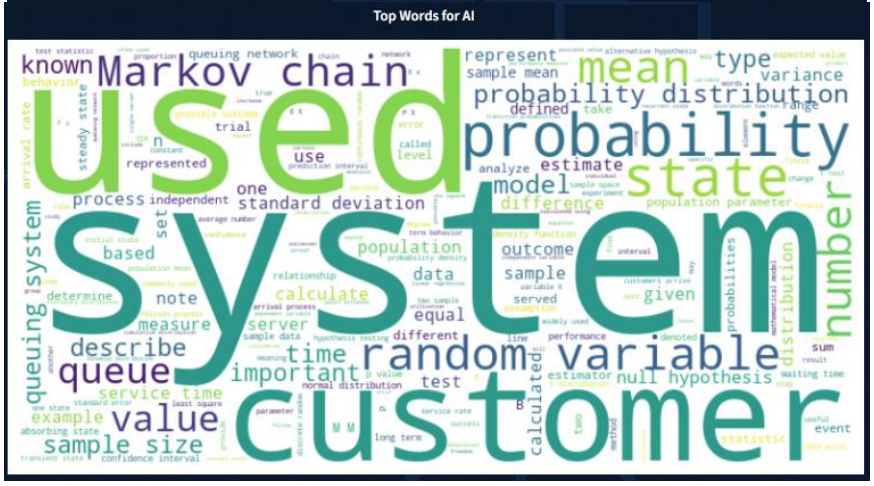


- Missing Values Heatmap: The heatmap visualizes missing values across three columns—**Question**, **Human**, and **AI**—with yellow indicating missing data, mostly concentrated in the **Human** and **AI** columns between row indices ~1560 to ~2000.

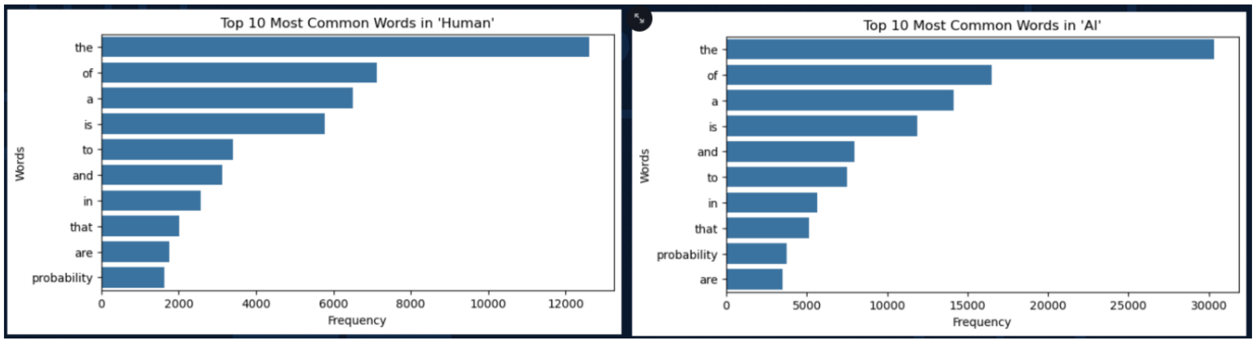


- Word Clouds: The word clouds show that both Humans and AI frequently used words like **"system," "state," "used," "customer,"** and **"probability,"** with slight variations in prominence—AI emphasizes **"used"** and **"customer"** more, while Humans highlight **"state"** and **"mean"** more often.

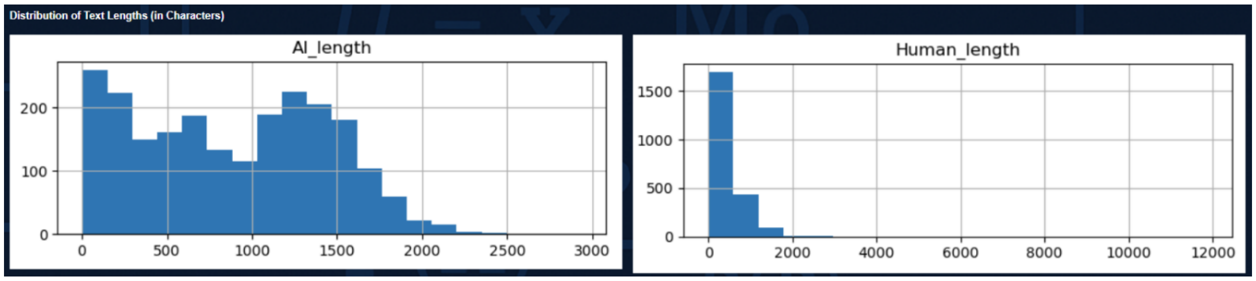




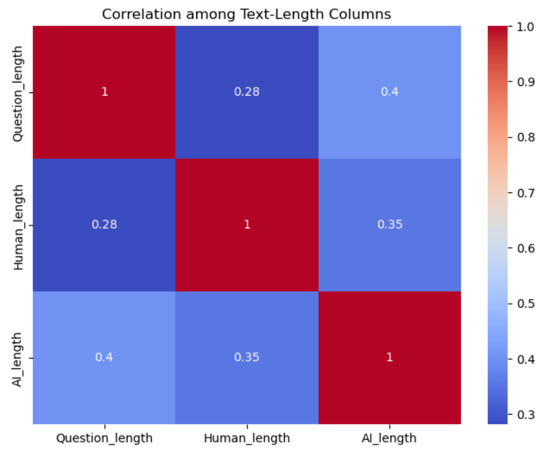
- Top 10 Most Common Words: The bar charts show that both Human and AI responses heavily rely on common stopwords like **"the," "of," "a," "is,"** and **"to,"** with **AI** using them significantly more frequently than **Humans**, particularly "the," which appears nearly **30,000 times** in AI responses compared to **13,000** in Human ones.



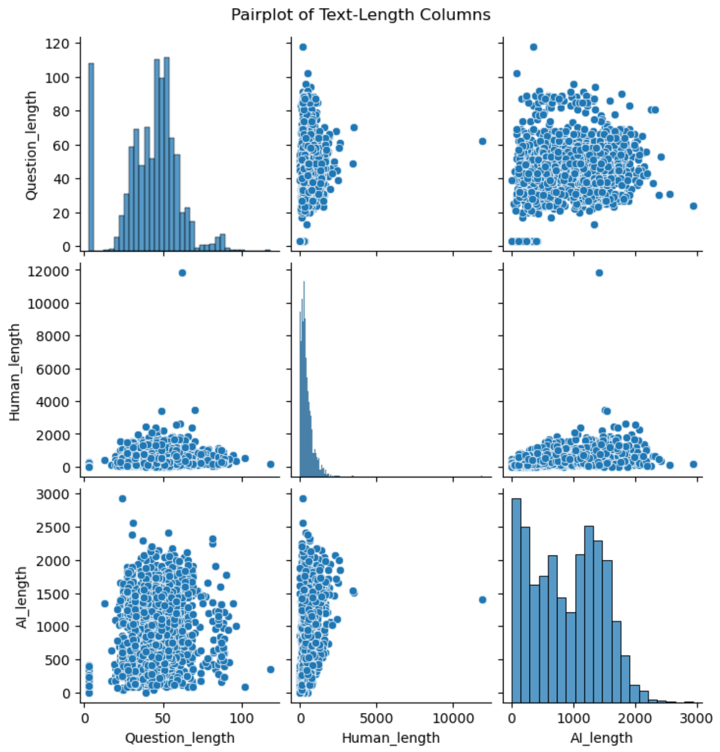
- Text Length Distributions: The histograms show that **AI responses are generally longer and more varied in length**, while **Human responses are shorter and more consistently below 1000 characters**, indicating more concise answers from humans.



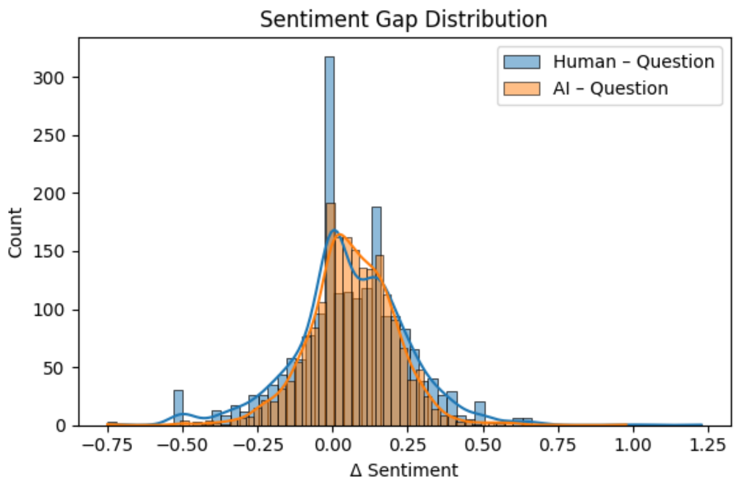
- Correlation Heatmap and Pairplot: The heatmap and pairplot reveal that **AI responses have a moderate correlation with question length (0.4)** and **human length (0.35)**, while **human responses show weaker correlations overall**, suggesting AI adjusts response length more consistently with input length than humans do.

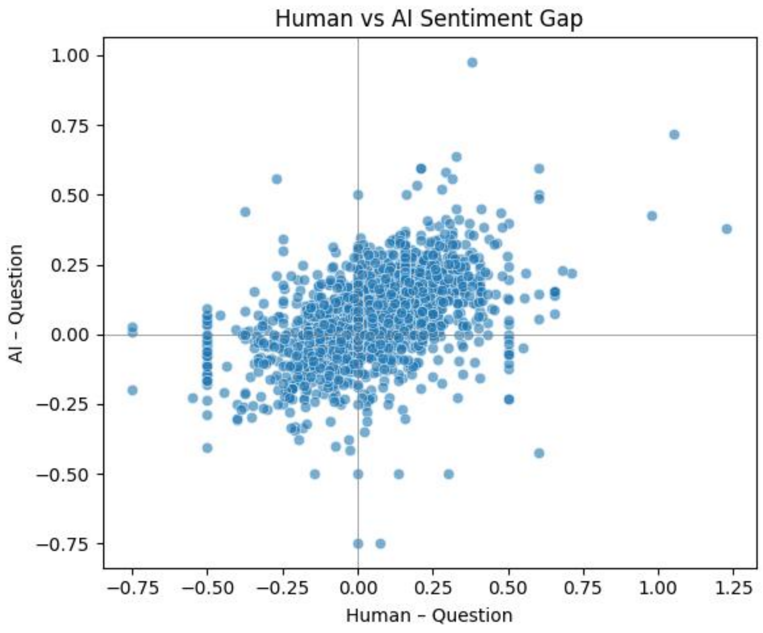


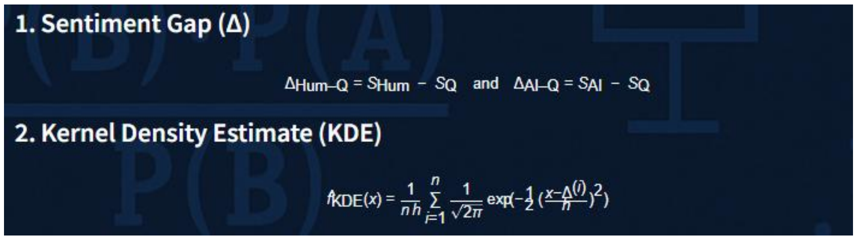
	Question_length	Human_length	AI_length
Question_length	1	0.28	0.4
Human_length	0.28	1	0.35
AI_length	0.4	0.35	1



- Sentiment Gap Analysis: The sentiment gap plots show that both AI and Human responses are generally close in sentiment to the questions (peaked near zero), but the scatterplot reveals a **moderate positive correlation** between their sentiment gaps—indicating that **when humans deviate in sentiment, AI tends to deviate similarly**.






1. Sentiment Gap (Δ)

$$\Delta_{Hum-Q} = S_{Hum} - S_Q \quad \text{and} \quad \Delta_{AI-Q} = S_{AI} - S_Q$$

2. Kernel Density Estimate (KDE)

$$\hat{f}_{KDE}(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\Delta^{(i)}}{h}\right)^2\right)$$

Sentiment Gap

$\Delta\_\text{Hum-Q} = S\_\text{Hum} - S\_Q$

$\Delta\_\text{AI-Q} = S\_\text{AI} - S\_Q$

Where:

- $\Delta\_\text{Hum-Q}$: Sentiment gap between the Human response and the Question
- $\Delta\_\text{AI-Q}$: Sentiment gap between the AI response and the Question
- S_Hum: Sentiment score of the Human response
- S_AI: Sentiment score of the AI response
- S_Q: Sentiment score of the original Question

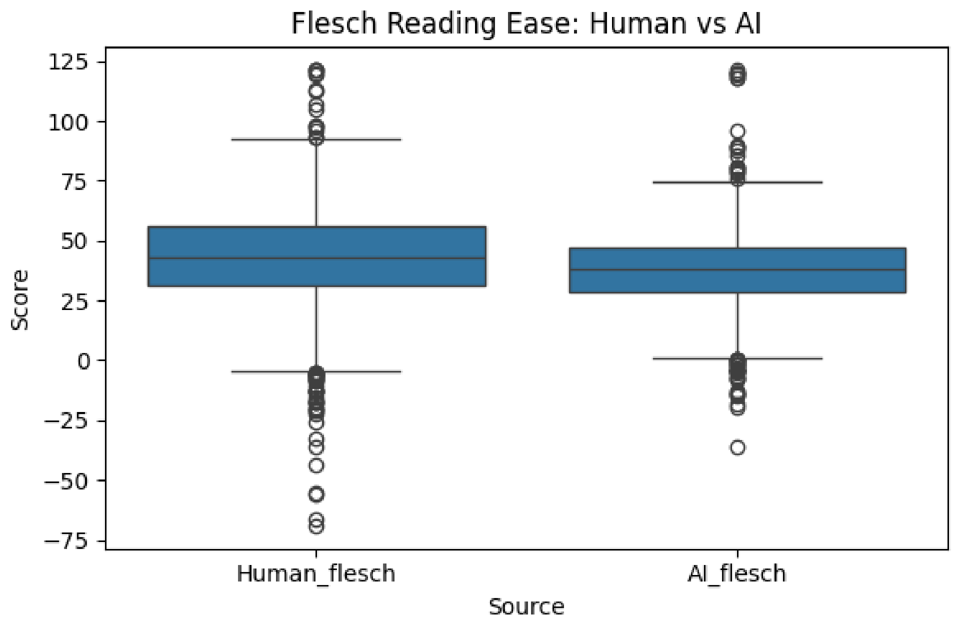
Kernel Density Estimate

$\hat{f}\_\text{KDE}(x) = (1 / nh) \sum (i=1 \text{ to } n) [1 / \sqrt{(2\pi)}] * \exp(-0.5 * ((x - \Delta\text{\^{}}(i)) / h)\text{\^{}}2)$

Where:

- $\hat{f}\_\text{KDE}(x)$: Estimated probability density function at point x
- n: Total number of data points (sentiment gap values)
- h: Bandwidth (smoothing parameter)
- $\Delta\text{\^{}}(i)$: The i-th observed sentiment gap
- x: Point at which the density is being estimated
- exp: Exponential function
- π: Mathematical constant Pi ≈ 3.1416

- Flesch Reading Ease: The boxplot shows that **Human responses generally have higher Flesch Reading Ease scores** than AI responses, indicating they are slightly easier to read on average, with both exhibiting similar variability and outliers.



Flesch Reading Ease (FRE)

$$\mathrm{FRE}(X) = 206.835 - 1.015\,\frac{W}{S} - 84.6\,\frac{\mathrm{SYL}}{W}$$

Gunning Fog Index (GFI)

$$\mathrm{GFI}(X) = 0.4\left(\frac{W}{S} + 100\,\frac{\mathrm{CW}}{W}\right)$$

Flesch Reading Ease

FRE(X) = 206.835 – 1.015 × (W / S) – 84.6 × (SYL / W)

Where:

- W: Total number of words
- S: Total number of sentences
- SYL: Total number of syllables
- FRE(X): Flesch Reading Ease score for text X

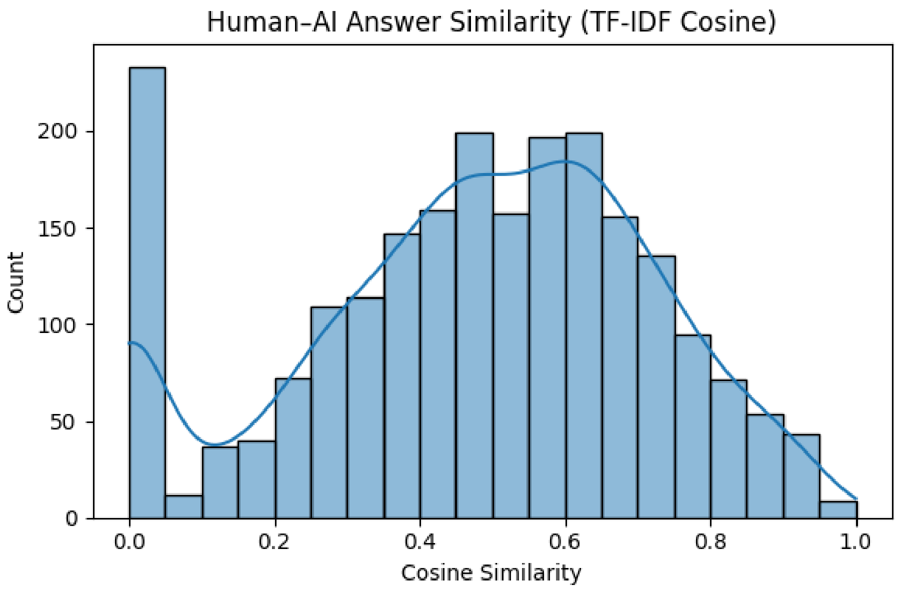
Gunning Fog Index

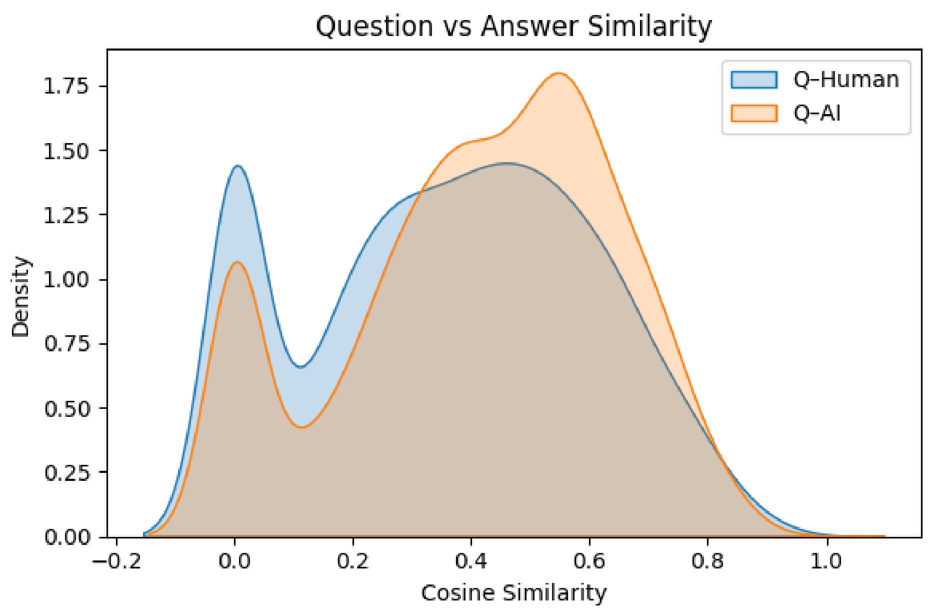
GFI(X) = 0.4 × ((W / S) + 100 × (CW / W))

Where:

- W: Total number of words
- S: Total number of sentences
- CW: Number of complex words (three or more syllables)
- GFI(X): Gunning Fog Index score for text X

- Text Similarity Analysis: The cosine similarity plots show that **AI responses tend to be more semantically aligned with the questions than Human responses**, with AI answers exhibiting higher average similarity scores to the questions, while Human-AI answer pairs vary more widely and often show lower or even zero similarity.





Modeling Approach

We evaluated:

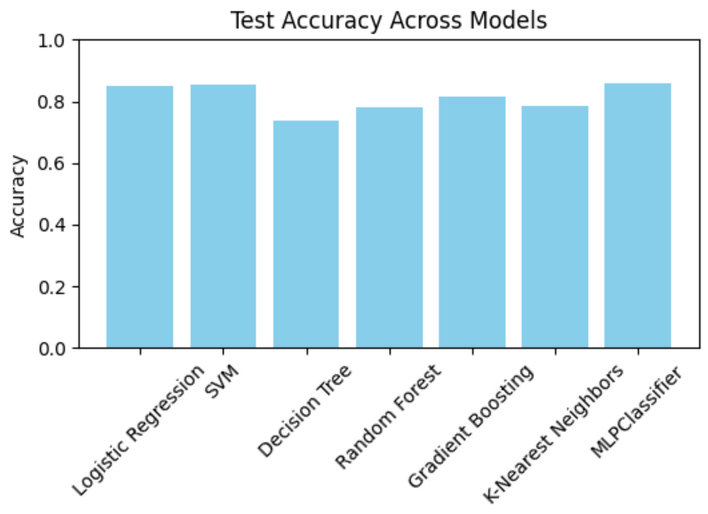
- Baseline Models: Logistic Regression, Linear SVM, Decision Tree, Random Forest, Gradient Boosting, K-Nearest Neighbors, MLP.
- Novel Model: BERT (pretrained and fine-tuned for binary classification).

Hyperparameter tuning was performed using 5-fold cross-validation.

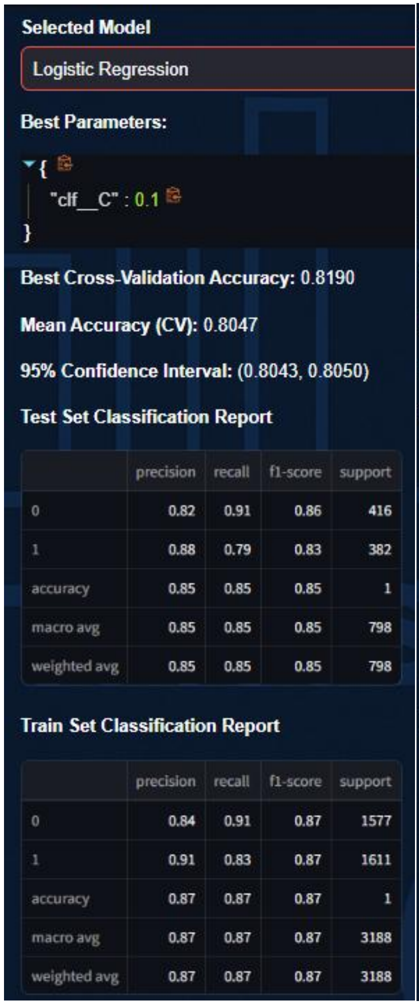
Traditional Classifier Performances:

Model Results

- Best Baseline: Logistic Regression, SVM and MLP achieved ~85% accuracy.



Logistic Regression, SVM, Decision Tree Performance:



Selected Model

Logistic Regression

Best Parameters:

```
{
  "clf__C" : 0.1
}
```

Best Cross-Validation Accuracy: 0.8190

Mean Accuracy (CV): 0.8047

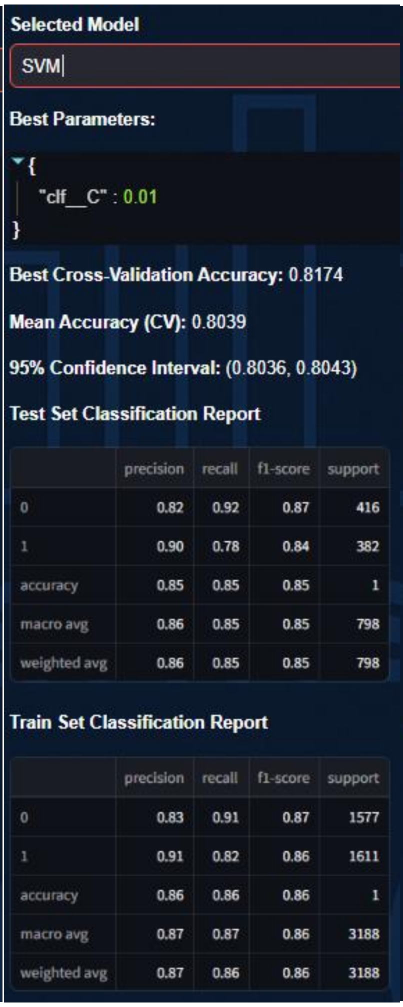
95% Confidence Interval: (0.8043, 0.8050)

Test Set Classification Report

	precision	recall	f1-score	support
0	0.82	0.91	0.86	416
1	0.88	0.79	0.83	382
accuracy	0.85	0.85	0.85	1
macro avg	0.85	0.85	0.85	798
weighted avg	0.85	0.85	0.85	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.84	0.91	0.87	1577
1	0.91	0.83	0.87	1611
accuracy	0.87	0.87	0.87	1
macro avg	0.87	0.87	0.87	3188
weighted avg	0.87	0.87	0.87	3188



Selected Model

SVM

Best Parameters:

```
{
  "clf__C" : 0.01
}
```

Best Cross-Validation Accuracy: 0.8174

Mean Accuracy (CV): 0.8039

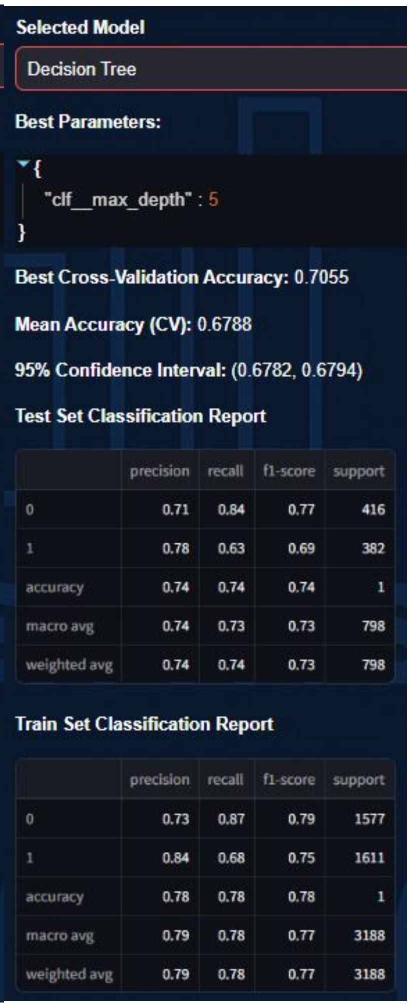
95% Confidence Interval: (0.8036, 0.8043)

Test Set Classification Report

	precision	recall	f1-score	support
0	0.82	0.92	0.87	416
1	0.90	0.78	0.84	382
accuracy	0.85	0.85	0.85	1
macro avg	0.86	0.85	0.85	798
weighted avg	0.86	0.85	0.85	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.83	0.91	0.87	1577
1	0.91	0.82	0.86	1611
accuracy	0.86	0.86	0.86	1
macro avg	0.87	0.87	0.86	3188
weighted avg	0.87	0.86	0.86	3188



Selected Model

Decision Tree

Best Parameters:

```
{
  "clf__max_depth" : 5
}
```

Best Cross-Validation Accuracy: 0.7055

Mean Accuracy (CV): 0.6788

95% Confidence Interval: (0.6782, 0.6794)

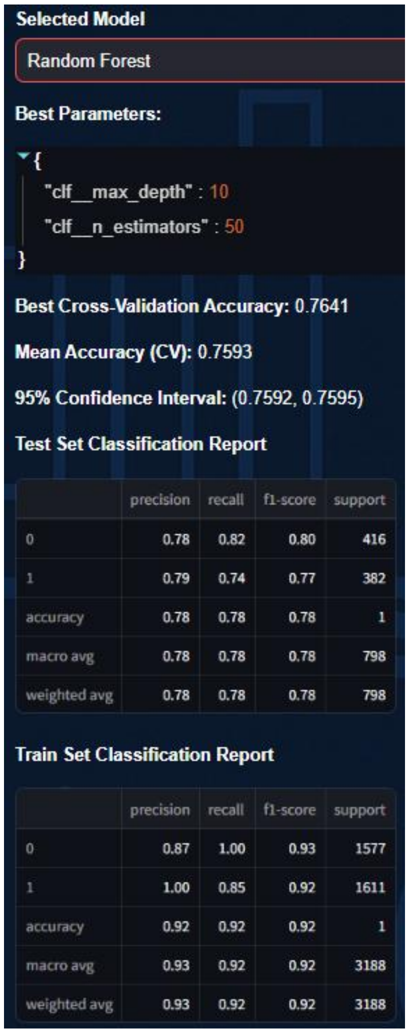
Test Set Classification Report

	precision	recall	f1-score	support
0	0.71	0.84	0.77	416
1	0.78	0.63	0.69	382
accuracy	0.74	0.74	0.74	1
macro avg	0.74	0.73	0.73	798
weighted avg	0.74	0.74	0.73	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.73	0.87	0.79	1577
1	0.84	0.68	0.75	1611
accuracy	0.78	0.78	0.78	1
macro avg	0.79	0.78	0.77	3188
weighted avg	0.79	0.78	0.77	3188

Random Forest, Gradient Boosting, K-Nearest Neighbors Performance:


Selected Model

Random Forest

Best Parameters:

```
{
  "clf__max_depth" : 10
  "clf__n_estimators" : 50
}
```

Best Cross-Validation Accuracy: 0.7641

Mean Accuracy (CV): 0.7593

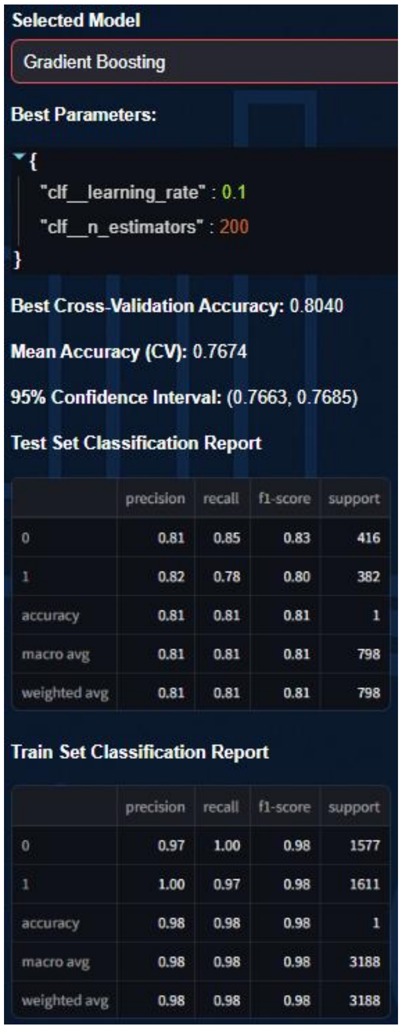
95% Confidence Interval: (0.7592, 0.7595)

Test Set Classification Report

	precision	recall	f1-score	support
0	0.78	0.82	0.80	416
1	0.79	0.74	0.77	382
accuracy	0.78	0.78	0.78	1
macro avg	0.78	0.78	0.78	798
weighted avg	0.78	0.78	0.78	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.87	1.00	0.93	1577
1	1.00	0.85	0.92	1611
accuracy	0.92	0.92	0.92	1
macro avg	0.93	0.92	0.92	3188
weighted avg	0.93	0.92	0.92	3188


Selected Model

Gradient Boosting

Best Parameters:

```
{
  "clf__learning_rate" : 0.1
  "clf__n_estimators" : 200
}
```

Best Cross-Validation Accuracy: 0.8040

Mean Accuracy (CV): 0.7674

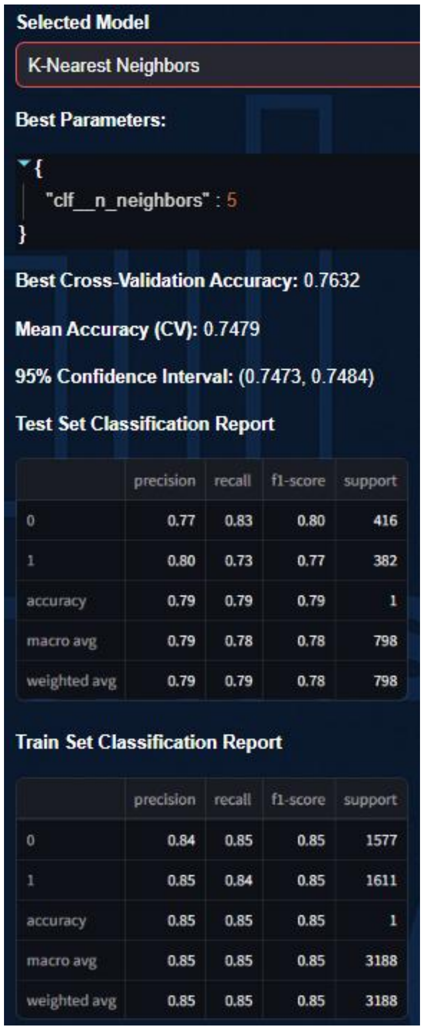
95% Confidence Interval: (0.7663, 0.7685)

Test Set Classification Report

	precision	recall	f1-score	support
0	0.81	0.85	0.83	416
1	0.82	0.78	0.80	382
accuracy	0.81	0.81	0.81	1
macro avg	0.81	0.81	0.81	798
weighted avg	0.81	0.81	0.81	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.97	1.00	0.98	1577
1	1.00	0.97	0.98	1611
accuracy	0.98	0.98	0.98	1
macro avg	0.98	0.98	0.98	3188
weighted avg	0.98	0.98	0.98	3188


Selected Model

K-Nearest Neighbors

Best Parameters:

```
{
  "clf__n_neighbors" : 5
}
```

Best Cross-Validation Accuracy: 0.7632

Mean Accuracy (CV): 0.7479

95% Confidence Interval: (0.7473, 0.7484)

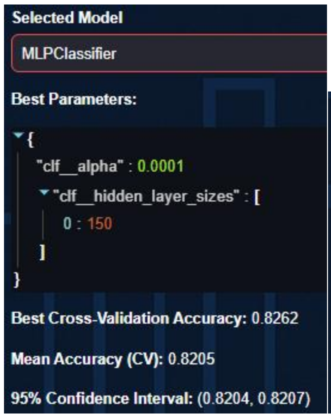
Test Set Classification Report

	precision	recall	f1-score	support
0	0.77	0.83	0.80	416
1	0.80	0.73	0.77	382
accuracy	0.79	0.79	0.79	1
macro avg	0.79	0.78	0.78	798
weighted avg	0.79	0.79	0.78	798

Train Set Classification Report

	precision	recall	f1-score	support
0	0.84	0.85	0.85	1577
1	0.85	0.84	0.85	1611
accuracy	0.85	0.85	0.85	1
macro avg	0.85	0.85	0.85	3188
weighted avg	0.85	0.85	0.85	3188

MLP Classifier Performance:


Selected Model

MLPClassifier

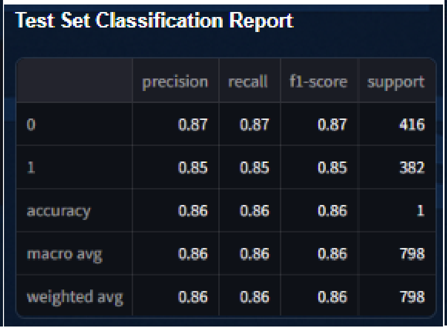
Best Parameters:

```
{
  "clf__alpha" : 0.0001
  "clf__hidden_layer_sizes" : [
    0 : 150
  ]
}
```

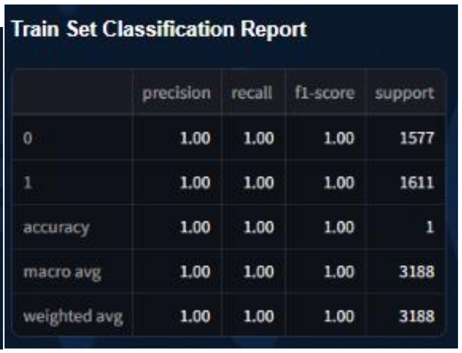
Best Cross-Validation Accuracy: 0.8262

Mean Accuracy (CV): 0.8205

95% Confidence Interval: (0.8204, 0.8207)


Test Set Classification Report

	precision	recall	f1-score	support
0	0.87	0.87	0.87	416
1	0.85	0.85	0.85	382
accuracy	0.86	0.86	0.86	1
macro avg	0.86	0.86	0.86	798
weighted avg	0.86	0.86	0.86	798


Train Set Classification Report

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1577
1	1.00	1.00	1.00	1611
accuracy	1.00	1.00	1.00	1
macro avg	1.00	1.00	1.00	3188
weighted avg	1.00	1.00	1.00	3188

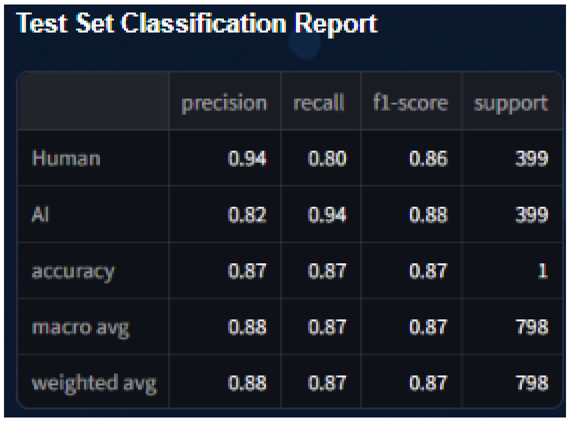
BERT Model Performance:

Tokenization using BERT Tokenizer
We use the BERT tokenizer from the "bert-base-uncased" model:

- Splits the input text into subword tokens (WordPiece tokenization).
- Converts tokens into corresponding token IDs that match BERT's vocabulary.
- Adds special tokens like [CLS] (for classification) and [SEP] (to separate segments).
- Pads or truncates each input to a fixed maximum length.
- Returns attention masks to distinguish real tokens from padding.

Model Training
We fine-tune a pretrained BERT model for sequence classification:

- Model: BertForSequenceClassification from Hugging Face.
- Optimizer: Typically AdamW for handling weight decay.
- Loss Function: Cross-entropy loss for classification tasks.
- Epochs: We train the model for 30 epochs, which is a good baseline for fine-tuning.
- Evaluation: During or after training, we track metrics like accuracy, precision, recall, and F1 score using a validation set.



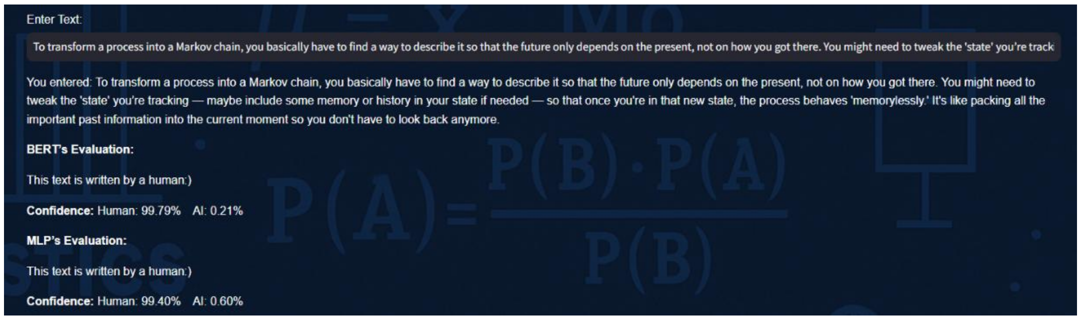
Test Set Classification Report

	precision	recall	f1-score	support
Human	0.94	0.80	0.86	399
AI	0.82	0.94	0.88	399
accuracy	0.87	0.87	0.87	1
macro avg	0.88	0.87	0.87	798
weighted avg	0.88	0.87	0.87	798

Working of the application –

Human Response -

To transform a process into a Markov chain, you basically have to find a way to describe it so that the future only depends on the present, not on how you got there. You might need to tweak the 'state' you're tracking — maybe include some memory or history in your state if needed — so that once you're in that new state, the process behaves 'memorylessly.' It's like packing all the important past information into the current moment so you don't have to look back anymore.



Enter Text:

To transform a process into a Markov chain, you basically have to find a way to describe it so that the future only depends on the present, not on how you got there. You might need to tweak the 'state' you're track

You entered: To transform a process into a Markov chain, you basically have to find a way to describe it so that the future only depends on the present, not on how you got there. You might need to tweak the 'state' you're tracking — maybe include some memory or history in your state if needed — so that once you're in that new state, the process behaves 'memorylessly.' It's like packing all the important past information into the current moment so you don't have to look back anymore.

BERT's Evaluation:

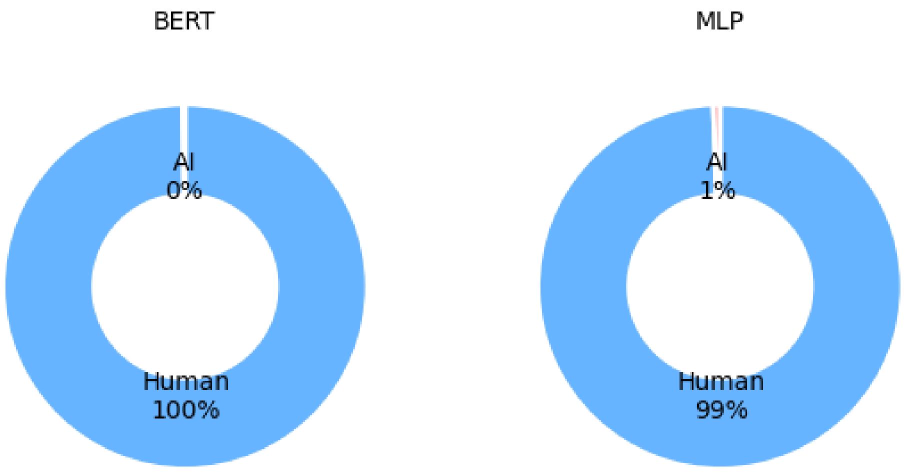
This text is written by a human:)

Confidence: Human: 99.79% AI: 0.21%

MLP's Evaluation:

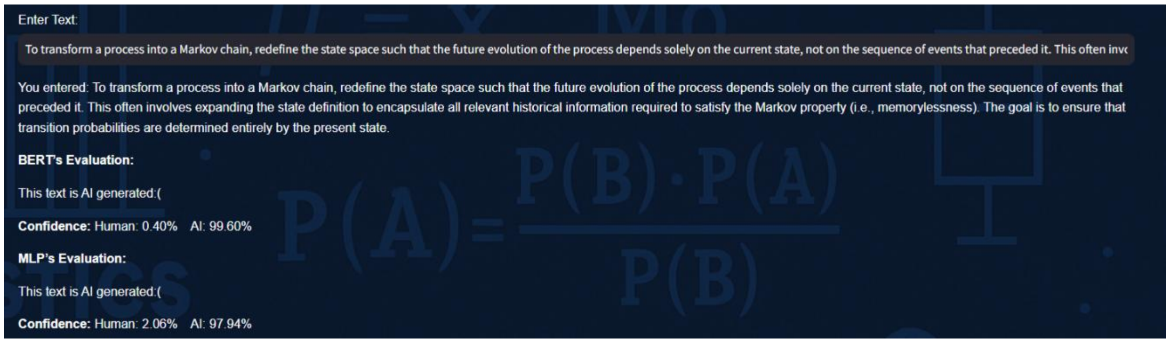
This text is written by a human:)

Confidence: Human: 99.40% AI: 0.60%



AI Response –

To transform a process into a Markov chain, redefine the state space such that the future evolution of the process depends solely on the current state, not on the sequence of events that preceded it. This often involves expanding the state definition to encapsulate all relevant historical information required to satisfy the Markov property (i.e., memorylessness). The goal is to ensure that transition probabilities are determined entirely by the present state.



Enter Text:

To transform a process into a Markov chain, redefine the state space such that the future evolution of the process depends solely on the current state, not on the sequence of events that preceded it. This often inv

You entered: To transform a process into a Markov chain, redefine the state space such that the future evolution of the process depends solely on the current state, not on the sequence of events that preceded it. This often involves expanding the state definition to encapsulate all relevant historical information required to satisfy the Markov property (i.e., memorylessness). The goal is to ensure that transition probabilities are determined entirely by the present state.

BERT's Evaluation:

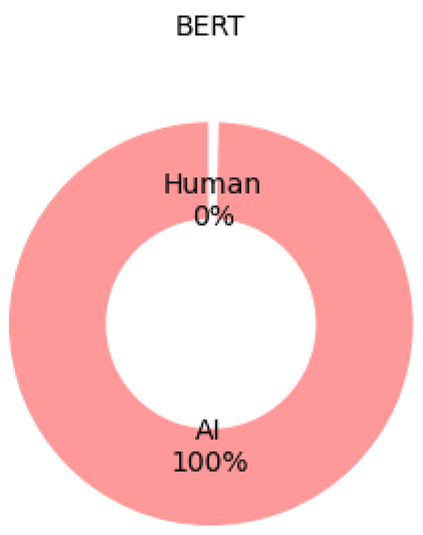
This text is AI generated:(

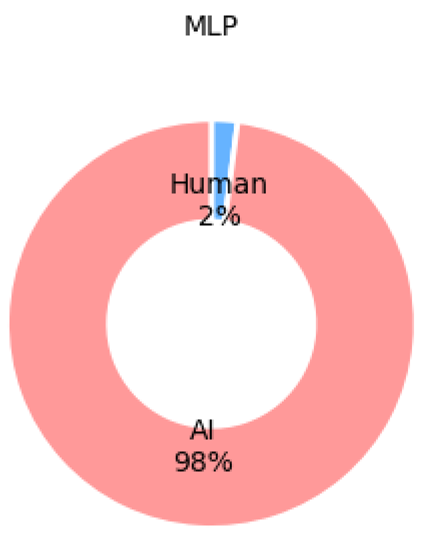
Confidence: Human: 0.40% AI: 99.60%

MLP's Evaluation:

This text is AI generated:(

Confidence: Human: 2.06% AI: 97.94%





Human Response

Insights and Discussion

- Human responses showed more variability and simpler phrasing.
- AI responses tended to be more formulaic and closer to the question text.
- Readability metrics and sentiment gaps proved important distinguishing features.
- Feature engineering, especially capturing text complexity and token patterns, was crucial for effective classification.

Conclusion

Our models, particularly BERT, can successfully distinguish between AI- and human-generated responses with high accuracy. This approach can help educational institutions uphold academic integrity. Future work could explore ensemble methods or more sophisticated feature engineering.

References

- **https://data.mendeley.com/datasets/mh892rksk2/4**
- **STT811 Stats Project GitHub Repository**
- **Project Streamlit App**
- **ChatGPT 4o assistance on concepts and code assistance.**