AI vs Human Text Detector

Presented by,

- Abirami Rajamanickam

Problem Statement

Objective:
Using a small dataset from Kaggle, Mathes is building a machine learning classifier to determine whether a given text was written by a human or generated by AI.

Challenge:
Distinguishing between human-written and AI-generated content with limited labeled data.

Source:
Kaggle Dataset – AI vs Human Text Classification Dataset

Data Collection

Data Source:

Kaggle Dataset

Collection Method:

Fetched using Kaggle’s Web API

Dataset Size:

Total Records: 1,299

Classes: AI-generated, Human-written

Format: CSV file with labeled text samples

Exploratory Data Analysis

Total Records: 1,299

Class Distribution: Balanced/Imbalanced analysis

Word Count Distribution: Compared across both classes

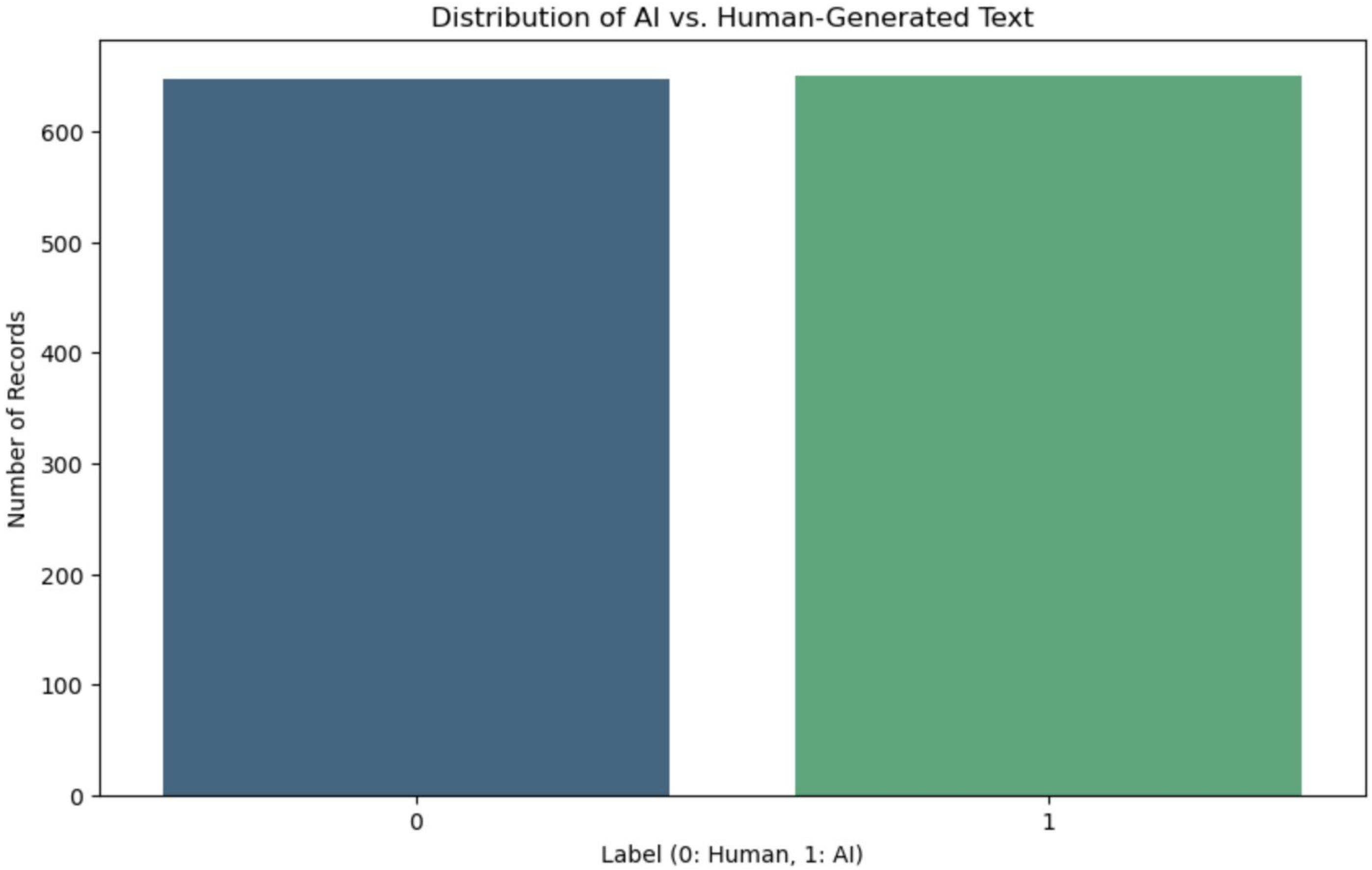
Top 5 Frequent Words: Visualized separately for AI and Human texts

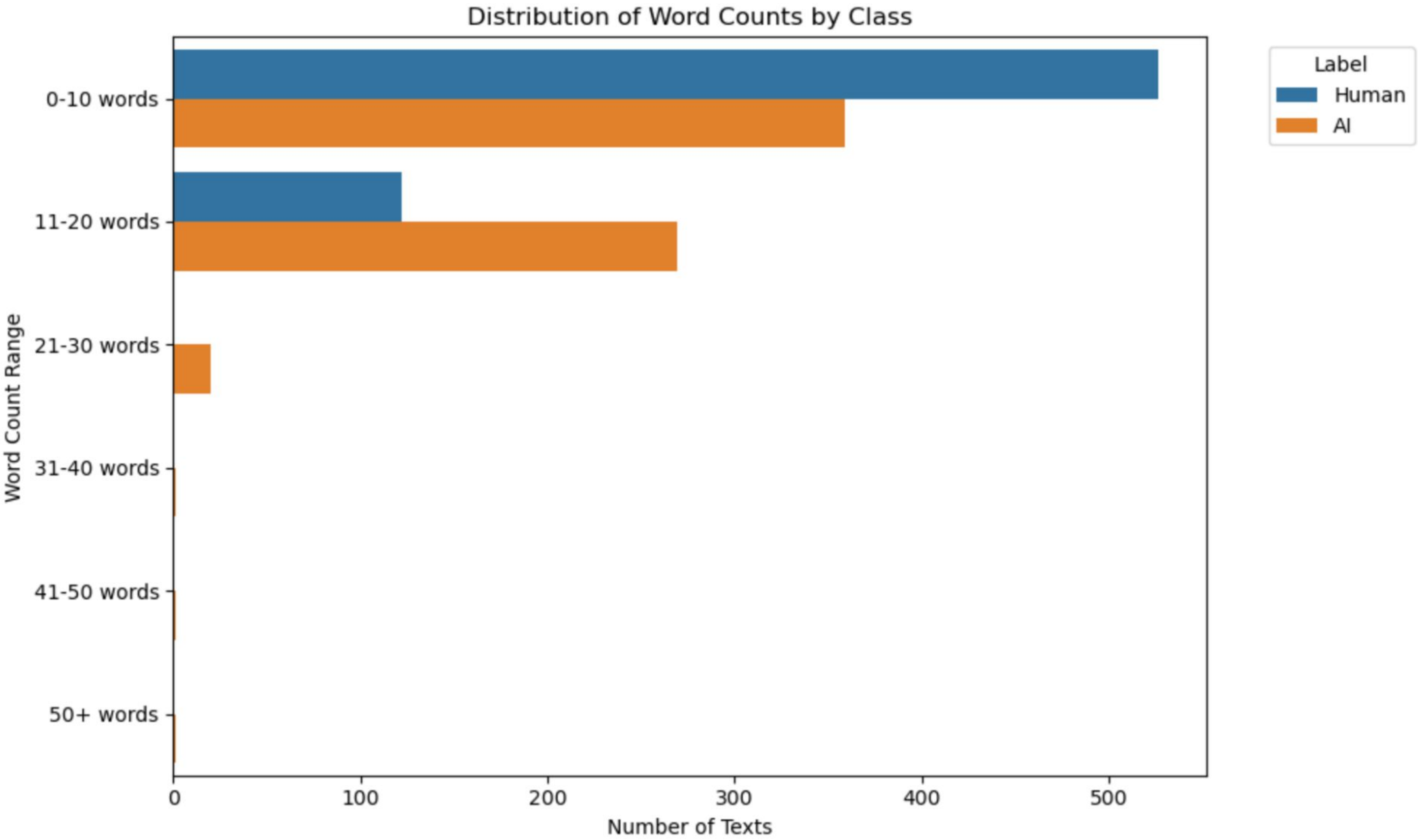
Visuals Included:

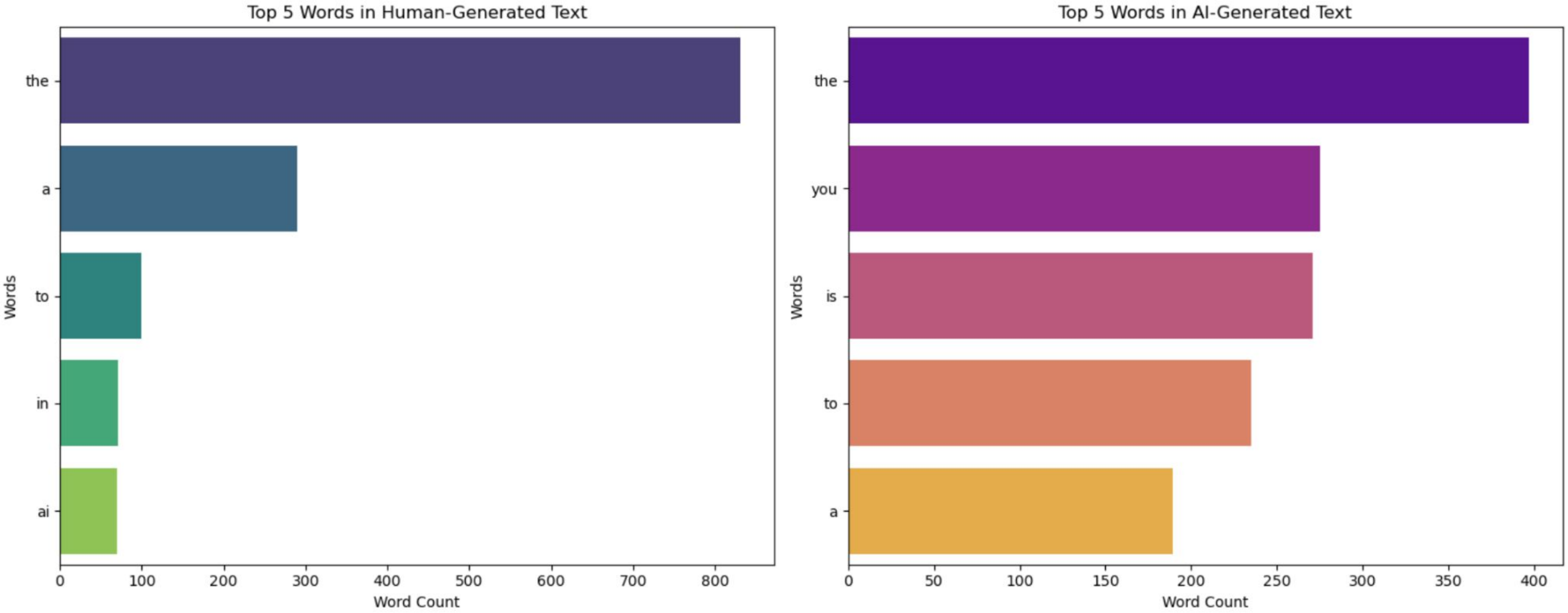
- Bar charts for class distribution - Even Class distribution
- Histogram for word count -

- Word clouds / bar charts for top terms

```
Counts of each class in the 'label' column:
label
1    651
0    648
Name: count, dtype: int64
```




Distribution of Word Counts by Class
Label
Human
AI
Word Count Range
0-10 words
11-20 words
21-30 words
31-40 words
41-50 words
50+ words
0
100
200
300
400
500
Number of Texts


Top 5 Words in Human-Generated Text
Words
the
a
to
in
ai
0
100
200
300
400
500
600
700
800
Word Count
Top 5 Words in AI-Generated Text
Words
the
you
is
to
a
0
50
100
150
200
250
300
350
400
Word Count

Data Modeling

Goal:

Binary classification – Predict whether input text is **AI-generated** or **Human-written**

Baseline Model: Why Logistic Regression?

Chosen for its simplicity and interpretability for Binary Classification problem

Achieved **95% accuracy** on training data

Served as benchmark for further model improvements

Model Evaluation

Final Model: Naive Bayes Classifier

Selected due to strong performance with text data and probabilistic outputs

Achieved **97% accuracy** on training data

Outperformed Logistic Regression

Why Naive Bayes?

Simple Model for Binary classification problem

Works well with small datasets

Model Performance and Score

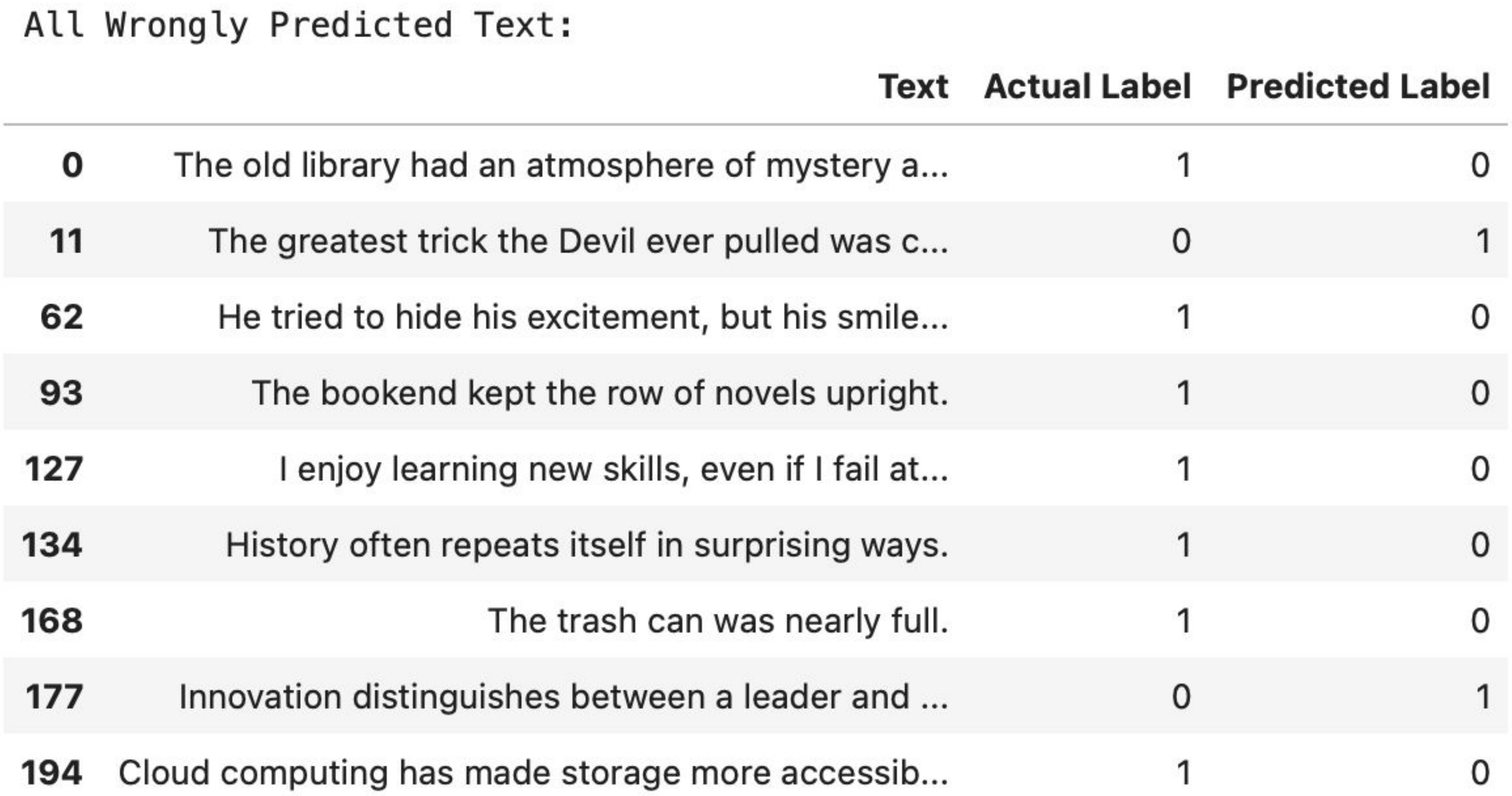
Performance Metrics:

Accuracy: 97%

Confusion Matrix: Visualized for error analysis

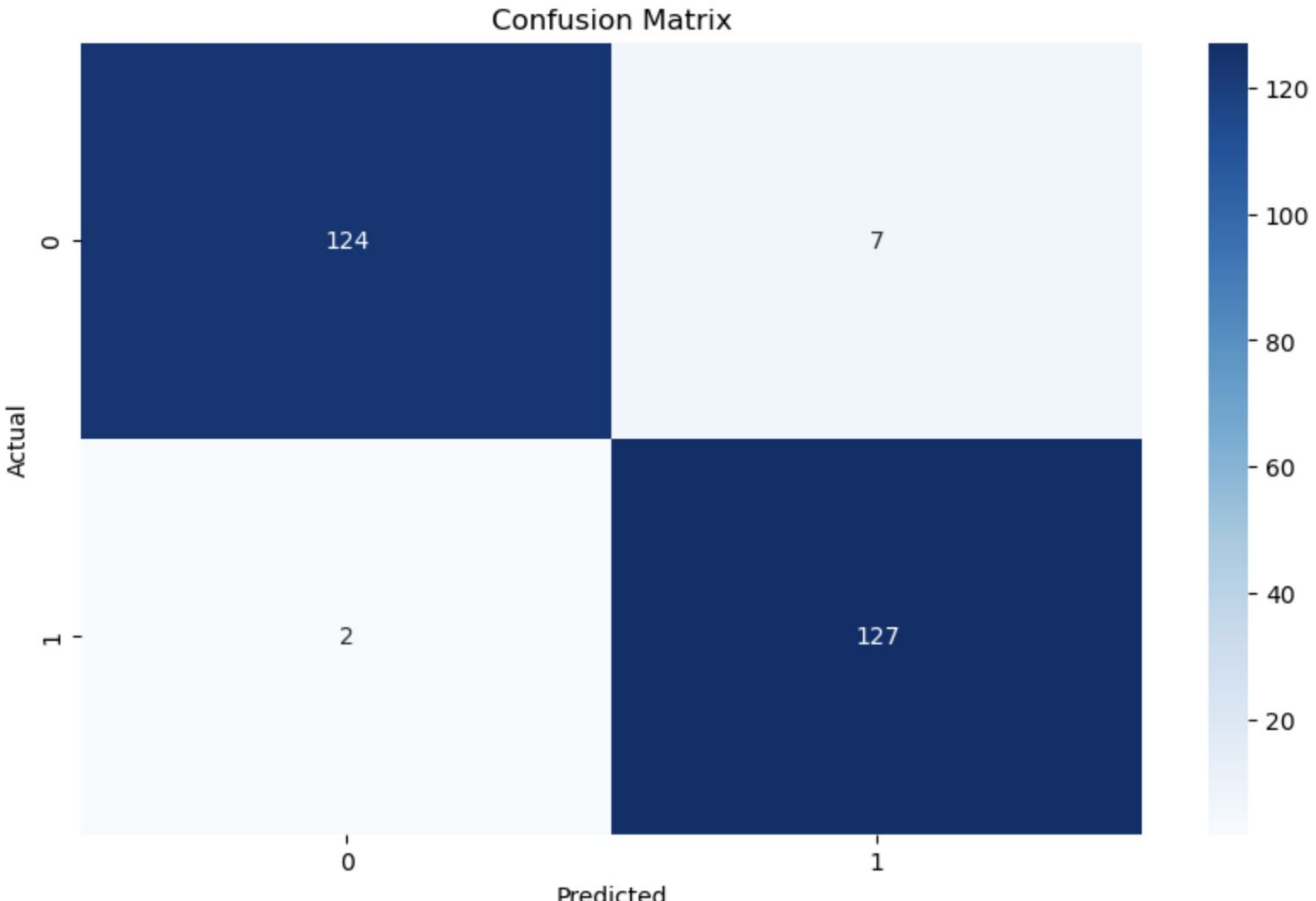
ROC Curve: Demonstrated classification threshold performance

Misclassified Examples: Highlighted for model interpretability



```
All Wrongly Predicted Text:
```

	Text	Actual Label	Predicted Label
0	The old library had an atmosphere of mystery a...	1	0
11	The greatest trick the Devil ever pulled was c...	0	1
62	He tried to hide his excitement, but his smile...	1	0
93	The bookend kept the row of novels upright.	1	0
127	I enjoy learning new skills, even if I fail at...	1	0
134	History often repeats itself in surprising ways.	1	0
168	The trash can was nearly full.	1	0
177	Innovation distinguishes between a leader and ...	0	1
194	Cloud computing has made storage more accessib...	1	0


Confusion Matrix
Actual
0
1
124
7
2
127
0
1
Predicted
120
100
80
60
40
20

Conclusion and Future works

Key Takeaways:

Naive Bayes provided best performance for limited data

High accuracy achieved with minimal preprocessing

The model did a great job of correctly identifying the two categories it was trained to recognize. Out of 260 total cases, it got 97% of them right. This means it made only a few mistakes.

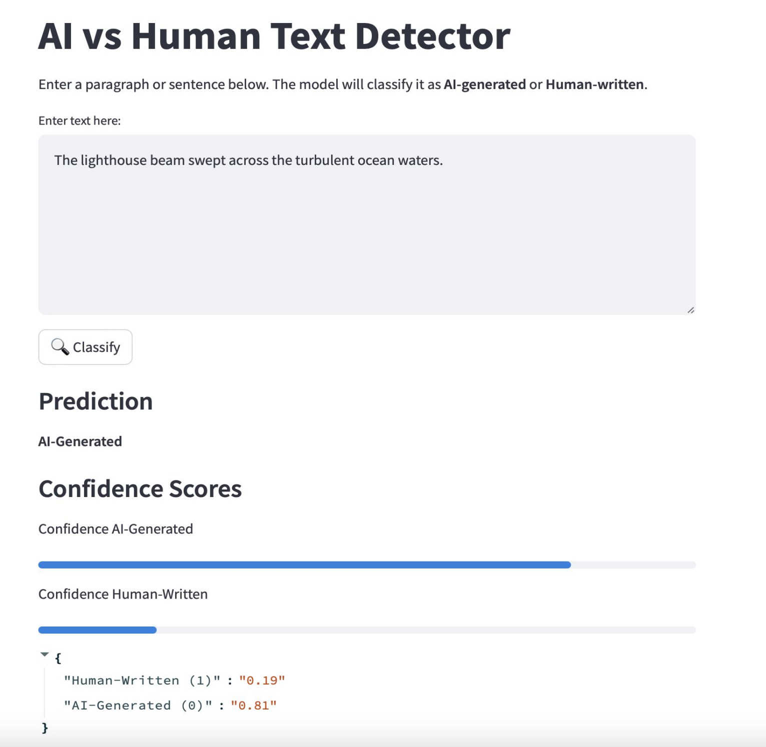
- It was especially good at identifying both types of cases, with only a small number of mix-ups.
- Overall, the model is accurate, reliable, and balanced in its decisions with the smaller dataset.

These results show that the model can be trusted to make correct predictions most of the time.

Future Work:

Expand dataset with more diverse samples

Streamlit Application Demo



AI vs Human Text Detector

Enter a paragraph or sentence below. The model will classify it as **AI-generated** or **Human-written**.

Enter text here:

The lighthouse beam swept across the turbulent ocean waters.

Prediction

AI-Generated

Confidence Scores

Confidence AI-Generated

Confidence Human-Written

```
{
  "Human-Written (1)" : "0.19"
  "AI-Generated (0)" : "0.81"
}
```

Thank you!