AI vs Human Text Detector

Objective:

Detect whether a given text sample is authored by a human or generated by an AI model using Natural Language Processing (NLP) techniques.

Use Cases:

- Identifying AI-generated academic submissions.

- Detecting fake product reviews written by bots/AI agents in E-commerce.

- Detecting and identifying AI-generated social media content in media and advertising.

Dataset source and structure:

Source: https://www.kaggle.com/datasets/prince7489/ai-vs-human-comparison-dataset/data

Data Description:

- **id**: ID of column
- **label**: If column is ai or human (Target variable)
- **topic**: Topic of the column
- **length_chars**: Number of characters in the text
- **length_words**: Total number of words in the text
- **quality_score**: Quality score of text
- **sentiment**: Sentiment score of the text
- **source_detail**: Source of the text
- **timestamp**: Timestamp of the text
- **plagiarism_score**: Plagiarism score of the text
- **notes**

The following analyses was done on the dataset containing both human text and AI generates text:

- **Data Cleaning & Lemmatization**: To remove whitespace and special characters and reduce words to their base form.

- **Most common words per class**: For identifying the most frequent words in AI and Human dataset.

- **Word Clouds for Human vs AI**: A visual representation and summary of a word, where the size represents its frequency.

- **N-Gram Analysis**: For analyzing sequences of N words, to capture writing style and phrases commonly use.

- **Complexity measures (Flesch score & lexical diversity)**: To capture complexity and readability of text.

- **Bag of words**: Counts every word in the text of each label

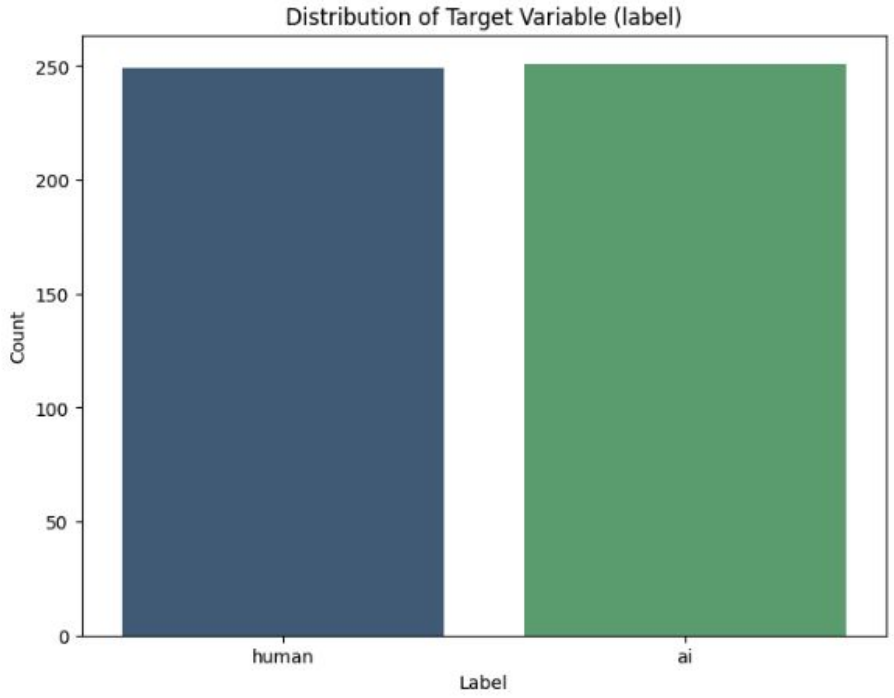
- **TF-IDF**: Weights words by how unique they are to a specific document compared to the whole dataset.

- **Sentence Embeddings**: Converts sentences into a vector for capturing semantic meaning.

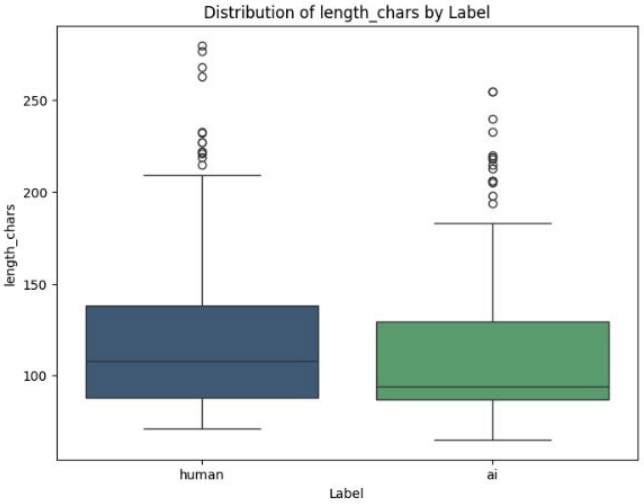
- **LLM-based embeddings:** Uses LLMs to capture semantic meaning and detects the AI signature of text.

- **SHAP & LIME Explainability:** Explains the most important feature that influenced the prediction result and how specific features influences the model's prediction.

Exploratory Data Analysis of Dataset



The dataset contained equal number of human and AI labeled data

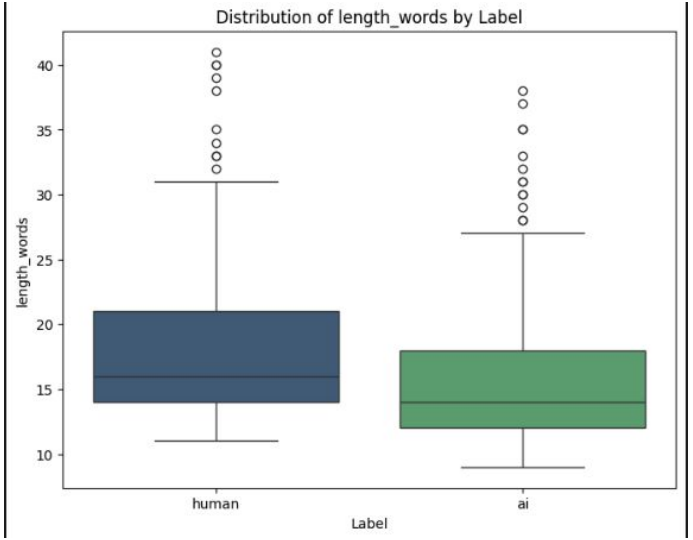


The box plot of length_chars vs label shows human text have a higher median length around approximately 110 characters, while AI text is slightly shorter with approximately 95 characters.

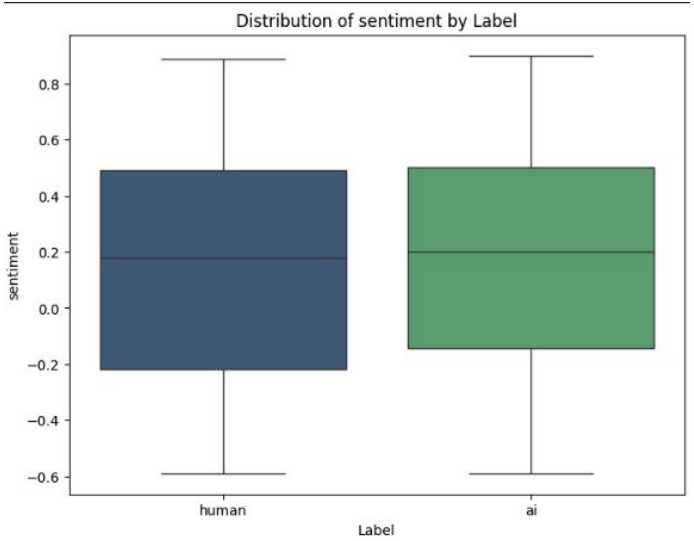
In summary, human text have longer characters and less predicted length while AI text have shorter characters and a more uniform length.

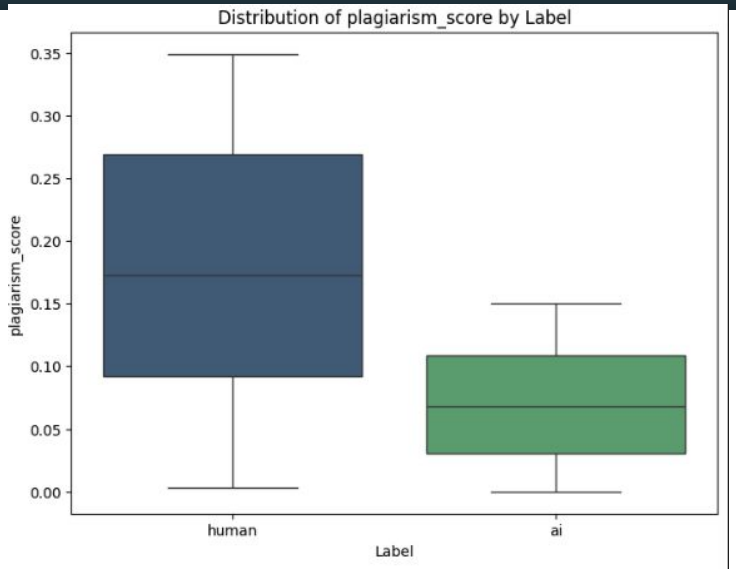
The box plot of length_words vs label shows human text have a higher median words length around approximately 16 words, while AI text is slightly shorter with approximately 14 words.

This shows human text have more words compared to AI text which are shorter and compact.

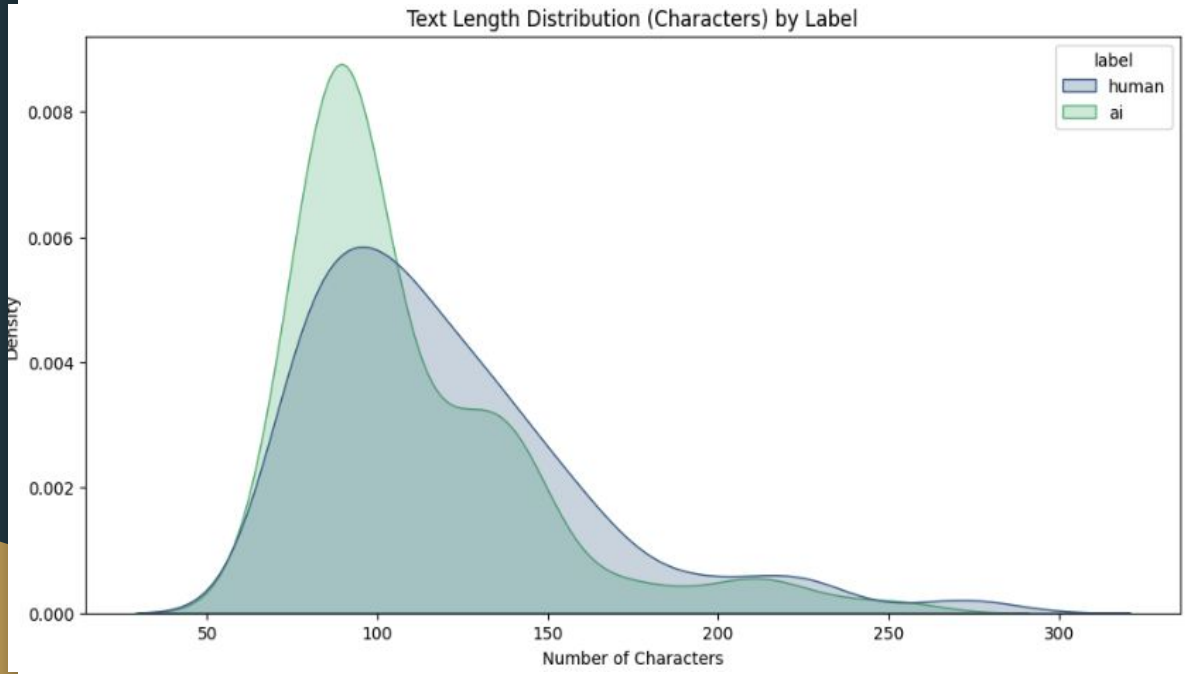


The box of plagiarism_score shows human text have a higher median plagiarism score of about 0.17, while AI have a much lower score of about 0.07. This indicates human have a higher chance of plagiarism.





The box of plagiarism_score shows human text have a higher median plagiarism score of about 0.17, while AI have a much lower score of about 0.07.

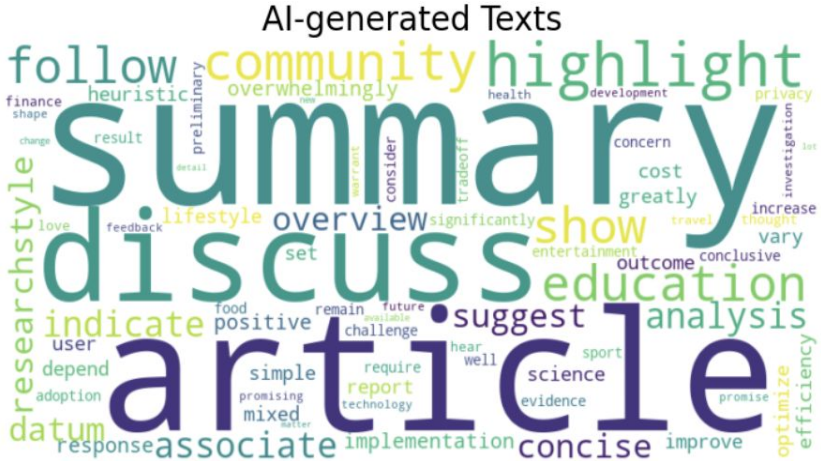


From the plot, the AI distribution is tighter and taller, with a peak around 90-100 characters. This suggests that the AI-generated text is consistent in it's length.

The human plot is broader with a peak around 100-110 characters, This indicates that human writing is more varied and less predictable than the AI output.

Exploratory Data Analysis of text

Word Clouds for Human vs AI Texts



Human texts

Includes more of personal life experiences words like **experience**, **believe**, **life**, **education**.

AI texts

Focuses on structured and analytical language like **summary**, **discuss** and **highlight** which suggests a systematic approach to information delivery

N-Gram Analysis

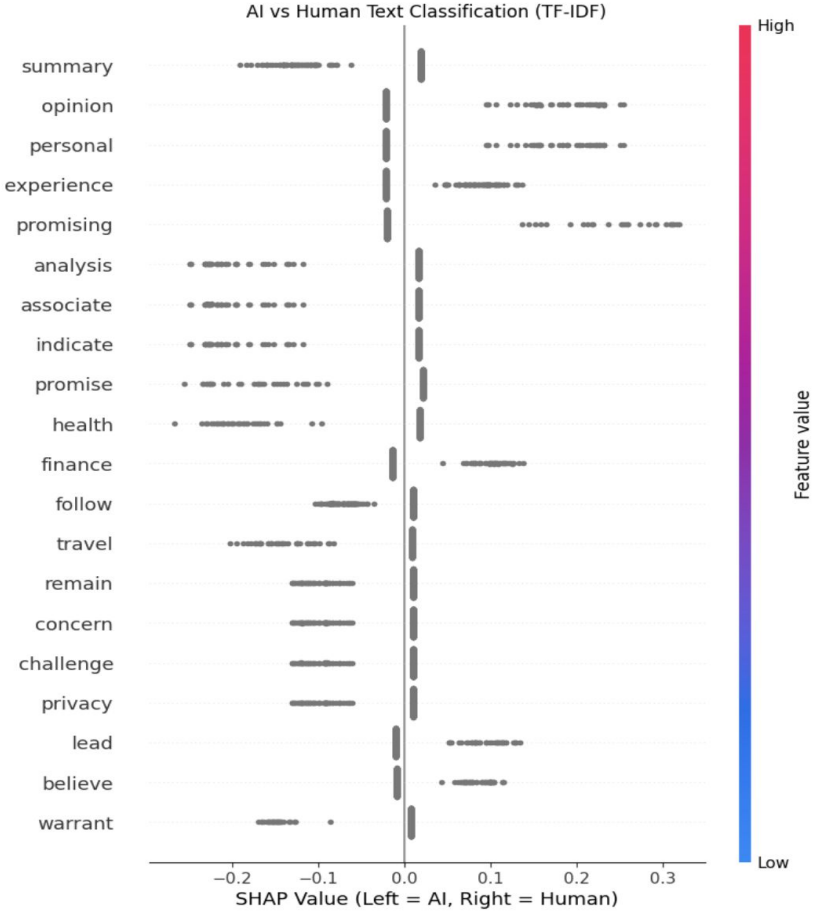
Top 2 words combinations for human text:

recently experience: 48 occurrence
daytoday life: 48 occurrence
personal opinion: 45 occurrence
try approach: 44 occurrence
approach relate: 44 occurrence
response overwhelmingly: 41 occurrence
overwhelmingly positive: 41 occurrence

Top 2 words combinations for AI text:

article discuss: 66 occurrence
follow summary: 43 occurrence
analysis indicate: 42 occurrence
researchstyle summary: 38 occurrence
concise overview: 37 occurrence
community response: 35 occurrence
optimize simple: 31 occurrence

SHAP Explainability



AI Indicators:
The diagram shows **associate**, **analysis**, **health** and **indicate** have negative SHAP values. This also suggest the model views these type of words as AI-generated content.

Human Indicators:
The diagram shows words like **personal**, **opinion** and **experience** have positive SHAP value and when these words appear, the model's confidence that the text is Human increases.

LIME Explainability

The diagram shows prediction result of a random dataset and how the result was gotten.

In this instance, the model shows probability of 67% AI and 33% human, words like "indicate", "analysis" and "associate" provided weight of 0.05 each, while words like "future", "shape" and "feedback" were seen as human generated words.