Results & Interpretation

Task 1: Project Repository Setup

A dedicated GitHub repository was established to host all project components, including code notebooks, data exploration outputs, cleaned datasets, and visualizations. This setup ensures version control, reproducibility, and transparent collaboration. Any updates to code or results are tracked automatically, enabling others to reproduce the experiments or build upon them.

Task 2: Dataset Loading and Preparation

The dataset *"Arabic Generated Abstracts"* was retrieved from Hugging Face using the datasets library in Google Colab. The dataset consists of three subsets:

1. **by_polishing** – AI-generated refinements of human-written abstracts.
2. **from_title** – AI-generated abstracts from paper titles alone.
3. **from_title_and_content** – AI-generated abstracts using both titles and paper content.

For analysis, data from the **by_polishing** subset was converted into a Pandas DataFrame. Each record contains a pair:

- A **human-written abstract** (original_abstract)
- An **AI-generated abstract** (allam_generated_abstract)

To facilitate comparative analysis, a combined dataset was created. Human abstracts were labeled as 0 and AI abstracts as 1. This transformation enabled supervised tasks such as classification and statistical comparison.

Task 3: Initial Data Exploration

a) Dataset Structure

Inspection of the dataset revealed a consistent schema across subsets, with two main text fields (human vs AI). This confirms the dataset's suitability for paired comparison tasks.

b) Distribution of Target Variable

Label distribution analysis demonstrated that the dataset is **balanced**: each human-written abstract has a corresponding AI-generated counterpart.

Interpretation: A balanced dataset reduces the risk of bias in downstream machine learning models, ensuring fairer classification between human and AI texts.

c) Data Quality Assessment

- **Missing values:** No significant missing entries were detected.
- **Duplicates:** Minimal duplication was observed, not enough to affect analysis.

Interpretation: The dataset is generally clean and requires little preprocessing, which enhances reliability for training and evaluation.

d) Exploratory Visualization

A histogram of **abstract word counts** was generated to compare human vs AI abstracts.

- Human abstracts exhibited broader variation in length, reflecting diverse writing styles.
- AI-generated abstracts tended to cluster more tightly, suggesting a more formulaic generation process.

Interpretation: While abstract length alone may not fully distinguish human from AI texts, it provides an early indicator of stylistic differences. More advanced linguistic and semantic features may further highlight divergences.

Summary of Findings

1. The dataset is **well-structured and balanced**, making it highly suitable for supervised learning tasks.
2. **High data quality** was confirmed, with negligible missing or duplicate entries.
3. Preliminary analysis revealed **length-based stylistic differences** between human and AI abstracts.

4. These results provide a strong foundation for subsequent steps, such as feature engineering, text similarity analysis, and classification modeling.