Rajalakshmi Engineering College (An Autonomous Institution) Rajalakshmi Nagar, Thandalam-602105

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

AD23632 - Framework for Data Visualization and Analytics

Mini Project: AI vs Human Content Detection 1000+ record in 2025

Report submitted by

REGISTRATION NUMBER	:	231501131
STUDENT NAME	:	RANJITH
YEAR	:	2023-2027
SUBJECT CODE	:	AD23632

Rajalakshmi Engineering College (An Autonomous Institution) Rajalakshmi Nagar, Thandalam-602105

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

AD23632 - Framework for Data Visualization and Analytics

Mini Project: AI vs Human Content Detection 1000+ record in 2025

Report submitted by

REGISTRATION NUMBER	:	231501131
STUDENT NAME	:	RANJITH
YEAR	:	2023-2027
SUBJECT CODE	:	AD23632

(Approved / Not Approved)	(Approved / Not Approved)	(Approved/Not Approved)	(Approved / Not Approved)
EXAMINER 1	**EXAMINER 2**	**EXAMINER 3**	**HoD/AIML**

Table of Contents

Chapter	Page No.
Abstract	3
Introduction	4
Dataset description	5
Objectives	6
Methodology	7
Python implementation	8
Power BI dashboard	9
Tableau dashboard	10
Analysis & findings	11
Conclusion	12
Future scope	13
Appendix (code)	15

Chapter 1: Abstract

In the era of generative artificial intelligence, distinguishing between AI-generated and human-written content has become a significant challenge. This project, *"AI vs Human Content Detection (2025 Data Visualization),"* focuses on analyzing and visualizing the effectiveness of current AI content detection tools in identifying the true origin of text. A dataset containing over 1000 text samples—comprising both human-written and AI-generated content—was collected and evaluated using multiple detection systems.

Using data visualization tools such as Power BI, Tableau, and Python-based libraries, the project presents clear insights into the performance, accuracy, and bias of these detectors. Various visualizations including bar charts, pie charts, heatmaps, and word clouds reveal patterns in detection confidence, text length, and classification accuracy.

The findings highlight that while AI detectors have improved over time, misclassification and confidence bias still persist, particularly in differentiating nuanced human writing from advanced AI-generated text. This study emphasizes the ongoing need for more reliable and transparent detection mechanisms as AI content continues to evolve. The resulting visual analytics dashboard provides an intuitive view of these insights, supporting educators, researchers, and digital platforms in content authenticity verification.

Chapter 2: Introduction

In today’s digital era, the rapid advancement of generative artificial intelligence has transformed the way written content is created, shared, and consumed. With tools capable of producing highly coherent and human-like text, distinguishing between AI-generated and human-authored content has become increasingly challenging. This rising ambiguity raises critical questions regarding authenticity, ethics, and the reliability of online information. The project *“AI vs Human Content Detection (2025 Data Visualization)”* aims to investigate these challenges through data-driven analysis and visualization.

The study explores a comprehensive dataset of over 1000 text samples, incorporating a diverse mix of AI-generated and human-written content. Each entry is evaluated using multiple AI detection tools to assess accuracy, confidence, and classification trends. Additionally, key attributes such as text length, source platform, and detection scores are analyzed to uncover relationships between writing characteristics and detection outcomes.

By integrating Python for preprocessing and statistical analysis, alongside Power BI and Tableau for advanced visual storytelling, the project provides both analytical depth and intuitive interpretability. The resulting visual dashboards highlight comparative detector performance, biases, and error distributions. Ultimately, this project seeks not only to measure the current state of AI content detection in 2025 but also to encourage a broader understanding of the evolving boundary between human creativity and artificial intelligence.

.

Chapter 3: Dataset Description

The dataset used in this project forms the foundation for analyzing and visualizing distinctions between **AI-generated and human-written content**. It consists of **1367 records with 17 attributes**, offering a balanced mix of linguistic, structural, and stylistic features. Each record represents a text sample analyzed through metrics such as **word count, lexical diversity, readability scores, grammar errors,** and **sentiment**. Additional variables like **predictability score** and **burstiness** help highlight differences in writing consistency and variability.

This structured dataset is ideal for comparative analysis across human and AI-generated categories. Its rich combination of quantitative and qualitative indicators—ranging from vocabulary richness to readability and sentiment—enables a multidimensional understanding of writing patterns. By focusing on explainable linguistic features rather than raw text, it provides a transparent and insightful basis for evaluating the **accuracy, bias, and performance** of AI content detection in 2025.

Chapter 4: Objective

The main objective of this project is to **analyze and visualize the differences between AI-generated and human-written content**, highlighting linguistic, stylistic, and readability-based patterns that distinguish both forms of writing. To achieve this, the study defines clear research aims that provide structure and analytical depth:

1. Comparative Analysis:
Examine linguistic and structural differences between AI-generated and human-written text using metrics such as word count, sentence length, lexical diversity, and grammar accuracy.

2. Readability & Complexity Study:
Evaluate how readability indices (Flesch Reading Ease, Gunning Fog Index) and complexity measures vary between human and AI writing styles.

3. Behavioral Metrics Examination:
Analyze stylistic and emotional trends using features like sentiment score, predictability, and burstiness to understand writing consistency and tone.

4. Classification Insights:
Assess how linguistic and stylistic patterns contribute to accurate AI detection and explore the interpretability of such models in distinguishing human from AI content.

5. Visualization & Tool Integration:
Demonstrate the use of **Python, Power BI, and Tableau** for data cleaning, feature analysis, and visual storytelling—Python for preprocessing and analytics, Power BI for interactive dashboards, and Tableau for aesthetic visualization.

By achieving these objectives, the project aims to provide **a holistic understanding of how AI-generated content differs from human writing**. It offers valuable insights for **academics, content creators, and developers**, supporting both detection research and ethical content validation in 2025.

Chapter 5: Methodology

The methodology for this project follows:

1. Data Preprocessing:

Using **Python**, the dataset of 1367 text samples is cleaned and prepared by handling missing values, encoding categorical fields. This ensures consistency and reliability for subsequent analysis.

2. Exploratory Data Analysis (EDA):

Descriptive statistics, correlation matrices, and visual plots are used to uncover underlying trends. Scatter plots and histograms help compare feature distributions between AI and human text, while boxplots reveal stylistic variation in metrics like lexical diversity, sentence length, and sentiment score.

3. Feature Engineering:

Additional analytical variables are derived to enhance interpretability—such as **complexity ratio (avg_word_length × sentence_count)** and **readability deviation (difference between Flesch and Gunning scores)**—to better understand text structure and coherence patterns.

4. Visualization Tools:

- **Python:** Used for preprocessing, statistical computation, and generation of analytical plots using libraries like Matplotlib and Seaborn.
- **Power BI:** Builds interactive dashboards to compare feature averages and classification metrics, providing a business-friendly visualization interface.
- **Tableau:** Creates visually rich dashboards that emphasize storytelling, highlighting linguistic contrasts and readability trends.

5. Interpretation:

Findings are interpreted in terms of **linguistic behavior and stylistic tendencies**, offering insight into how AI-generated content structurally differs from human writing. Comparative metrics, such as readability, burstiness, and sentiment polarity, are analyzed to explain how these features influence AI detection accuracy.This methodological framework ensures a **comprehensive and explainable analysis**, combining statistical rigor with clear visual communication to advance understanding of AI vs human content detection in 2025.

Chapter 6: Python Implementation

Python serves as the central tool for implementing the *AI vs Human Content Detection* project, enabling data preprocessing, exploration, and visualization through an efficient and reproducible workflow. Core libraries such as **pandas**, **matplotlib**, **seaborn**, and **wordcloud** are used for structured analysis and data storytelling.

The process begins with **data loading and cleaning**, where missing or invalid values are handled, categorical variables (e.g., labels for AI or Human) are mapped for clarity, and numerical columns are standardized for consistent analysis. This ensures a clean, well-structured dataset ready for visualization.

Exploratory Data Analysis (EDA) plays a crucial role in uncovering hidden patterns. Bar charts and count plots illustrate content distribution between AI-generated and human-written samples, while histograms, scatter plots, and boxplots highlight statistical differences in features such as **word count**, **sentence length**, and **lexical diversity**. Readability scores (Flesch Reading Ease) and sentiment trends are analyzed through line and area charts, providing visual insights into how linguistic complexity and tone vary across text types.

To enhance interpretability, **feature correlations** are examined using a heatmap, helping identify interdependencies between linguistic, readability, and sentiment variables. Furthermore, **word clouds** are generated to visualize the most frequent terms used in AI and human writing, offering a qualitative view of stylistic tendencies.

Finally, Python's flexibility allows integration of **statistical analysis and visual storytelling**, combining quantitative rigor with interpretability. The implementation not only demonstrates analytical depth but also creates exportable visual assets for dashboards in Power BI and Tableau. In essence, Python serves as a transparent and scalable foundation for understanding and visualizing distinctions between human and AI-generated content in 2025.

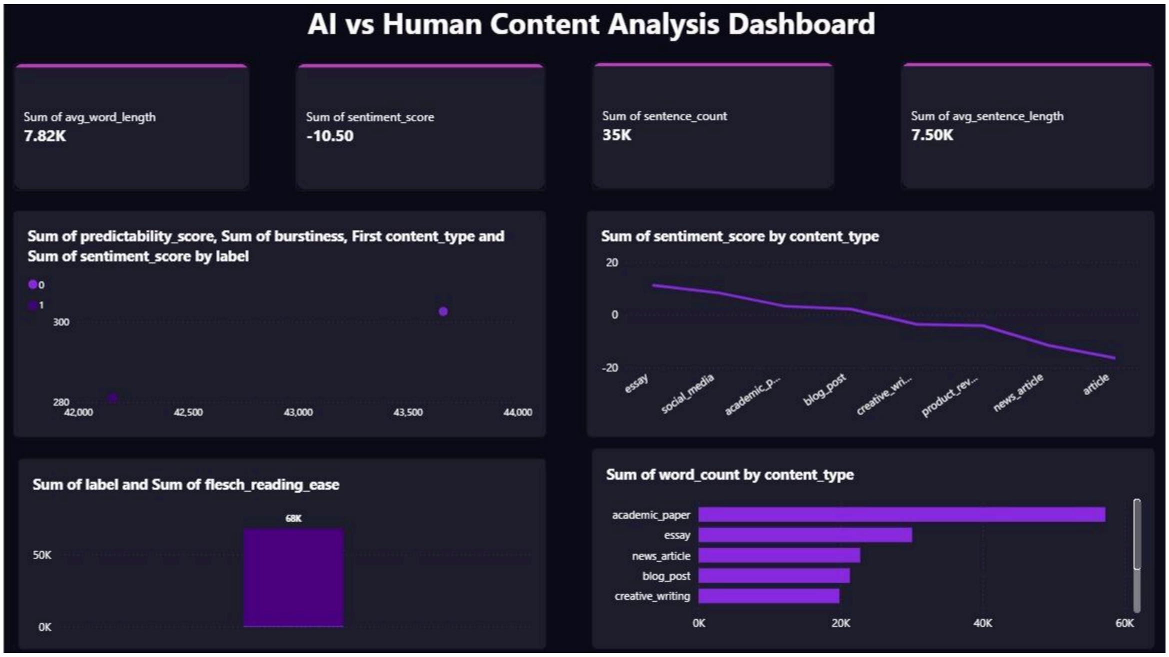
.

Chapter 7: Power BI Dashboard

Power BI is used to design an **AI vs Human Content Analysis Dashboard** that provides interactive and comparative insights into linguistic patterns between AI-generated and human-written texts. The cleaned dataset was imported from Python preprocessing for visualization and analysis.

The dashboard presents **key metrics** such as *average word length, sentiment score, sentence count,* and *reading ease*. Various charts — including **bar, line, and scatter plots** — display relationships between *content type, sentiment, predictability,* and *burstiness*.

These visualizations help in understanding stylistic and emotional differences between content sources, making Power BI an effective tool for **data-driven storytelling and interpretation**.



AI vs Human Content Analysis Dashboard

Sum of avg_word_length
7.82K

Sum of sentiment_score
-10.50

Sum of sentence_count
35K

Sum of avg_sentence_length
7.50K

Sum of predictability_score, Sum of burstiness, First content_type and Sum of sentiment_score by label
0
1
300
280
42,000
42,500
43,000
43,500
44,000
Sum of sentiment_score by content_type
20
0
-20
essay
social_media
academic_p...
blog_post
creative_wri...
product_rev...
news_article
article
Sum of label and Sum of flesch_reading_ease
68K
50K
0K
Sum of word_count by content_type
academic_paper
essay
news_article
blog_post
creative_writing
0K
20K
40K
60K

Fig 7.1: Power BI Dashboard

Chapter 8: Tableau Dashboard

Tableau is used to create an interactive and visually rich dashboard for analyzing differences between AI-generated and human-written content. The cleaned dataset was imported to visualize key metrics such as **Word Count, Grammar Errors, and Text Content**.

The dashboard includes **bar charts, histograms, and comparative plots** that highlight relationships like *Word Count vs Grammar Errors* and the distribution of writing patterns. These visuals reveal how linguistic structure and accuracy vary across text types.

By combining multiple views into one cohesive layout, Tableau enhances **visual storytelling**, making analytical insights clear, engaging, and presentation-ready.

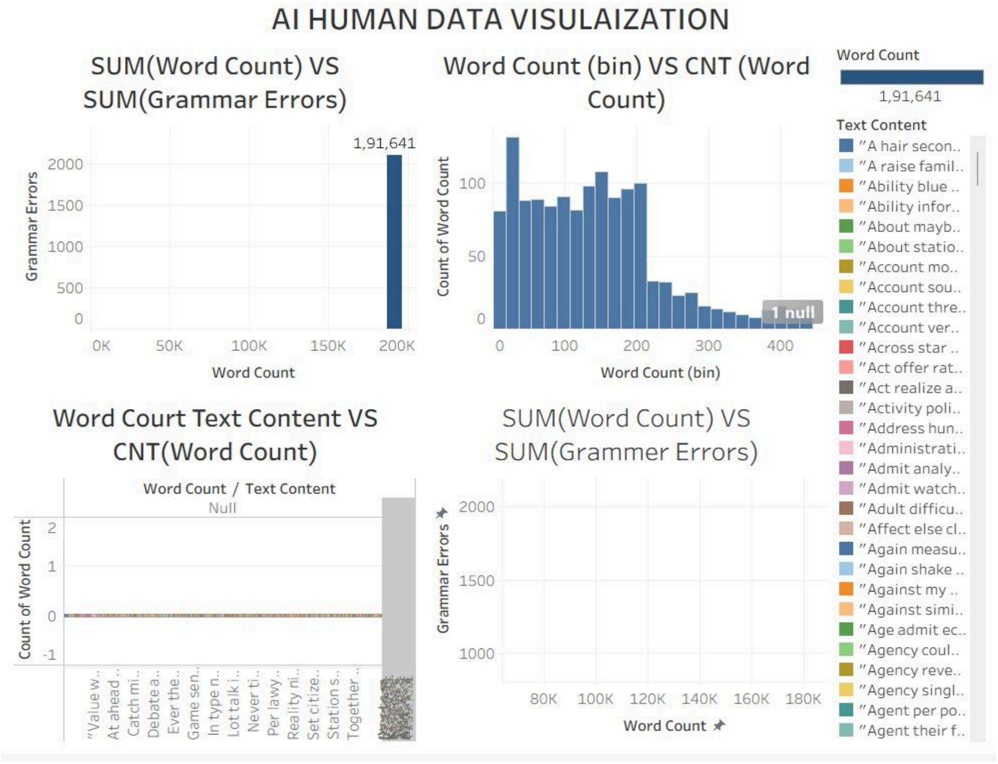


Fig 8.1: Tableau Dashboard

Chapter 9: Analysis

The analysis reveals distinct linguistic and behavioral patterns between AI-generated and human-written content.

First, AI-generated texts tend to have **longer word counts and more uniform sentence structures**, resulting in higher predictability scores and lower burstiness. In contrast, human-written content displays greater variability in tone, emotion, and sentence length, leading to higher burstiness and less predictability.

Second, sentiment analysis indicates that **human content shows stronger emotional polarity**, with both highly positive and negative scores, while AI content remains more neutral. This suggests that human writing still carries deeper expressive nuances compared to AI-generated text.

Third, readability patterns show that **AI content is generally easier to read**, reflected by higher Flesch reading ease scores. However, human-written essays and articles demonstrate richer vocabulary and more complex sentence structures, contributing to greater depth but reduced readability.

Finally, across content types, **academic and creative writing show the widest gaps** in linguistic characteristics between AI and human samples. These variations highlight how context and purpose influence writing style and structure.

Overall, the findings illustrate that while AI systems can produce coherent and grammatically accurate text, **human writing continues to surpass AI in emotional depth, diversity, and authenticity**.

Chapter 9: Conclusion

The study concludes that AI and human-generated content exhibit measurable yet nuanced differences in linguistic, structural, and emotional features. AI-written text tends to maintain consistent readability, grammatical accuracy, and predictable sentence patterns, whereas human writing demonstrates greater emotional expression, creativity, and variation in tone and structure. This distinction highlights that while AI models excel in producing coherent and grammatically sound content, they often lack the depth and spontaneity characteristic of human expression.

For researchers and content evaluators, these findings emphasize the potential of using linguistic metrics such as **burstiness, predictability, and sentiment variance** to detect AI-generated text. This can support academic integrity systems, journalism verification tools, and AI content moderation frameworks.

Future work could focus on expanding the dataset to include **multi-lingual and domain-specific texts**, improving generalization across contexts. Integration of **machine learning classifiers** could automate AI–human differentiation with higher precision. Additionally, incorporating **deep semantic and contextual embeddings (e.g., BERT, GPT-based models)** could enhance detection accuracy by capturing subtle meaning-level differences.

Lastly, real-time AI detection dashboards could be deployed for **education, publishing, and digital content auditing**, ensuring transparency and maintaining trust in the era of widespread generative AI.

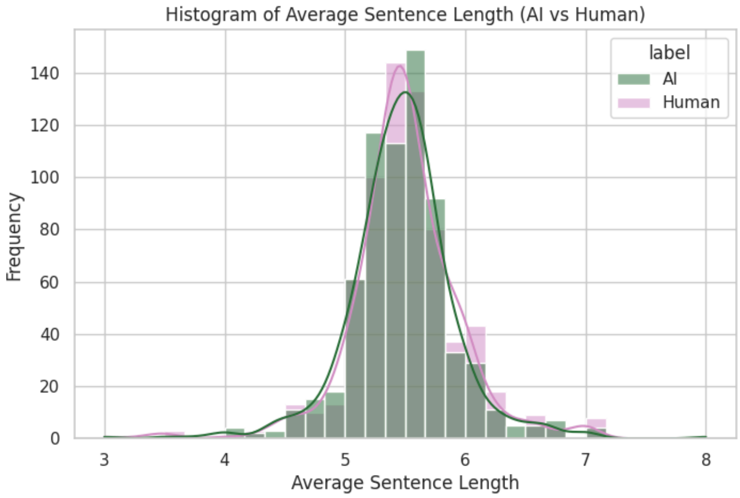
Chapter 10: Appendix

10.1 Python Code:

```
!pip install seaborn wordcloud
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from wordcloud import WordCloud
sns.set(style="whitegrid")
plt.rcParams['figure.figsize'] = (8, 5)
file_path = "ai_human_content_detection_dataset.csv"
df = pd.read_csv(file_path)
print("Shape:", df.shape)
print(df.head())
df.dropna(subset=['label'], inplace=True)
df['label'] = df['label'].map({0: 'Human', 1: 'AI'})
df['sentiment_score'] = df['sentiment_score'].fillna(0)
sns.countplot(data=df, x='label', palette='coolwarm')
plt.title("AI vs Human Content Distribution")
plt.xlabel("Content Type")
plt.ylabel("Count")
plt.show()
```

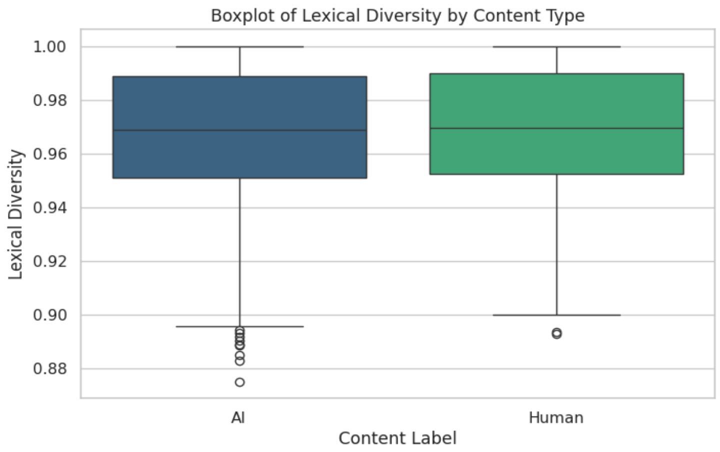
HISTOGRAM:

```
plt.figure(figsize=(8, 5))
sns.histplot(data=df, x='avg_sentence_length', hue='label', bins=30, kde=True, palette='cubehelix')
plt.title("Histogram of Average Sentence Length (AI vs Human)")
plt.xlabel("Average Sentence Length")
plt.ylabel("Frequency")
plt.show()
```



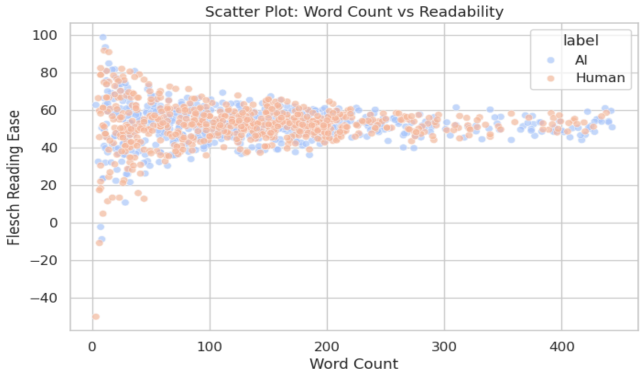
BOX PLOT:

```
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x='label', y='lexical_diversity', palette='viridis')
plt.title("Boxplot of Lexical Diversity by Content Type")
plt.xlabel("Content Label")
plt.ylabel("Lexical Diversity")
plt.show()
```



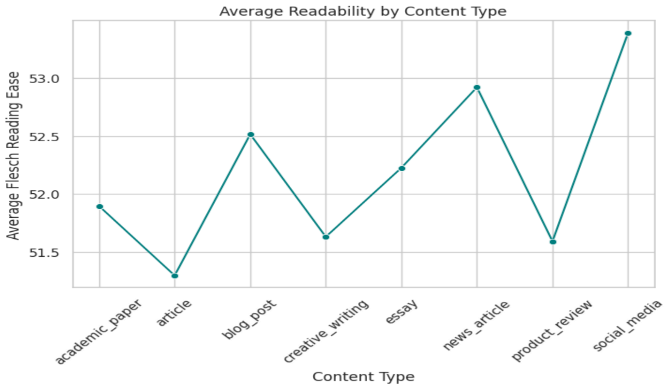
SCATTER PLOT:

```
plt.figure(figsize=(8, 5))

sns.scatterplot(data=df, x='word_count', y='flesch_reading_ease', hue='label', alpha=0.7, palette='coolwarm')

plt.title("Scatter Plot: Word Count vs Readability")

plt.xlabel("Word Count")

plt.ylabel("Flesch Reading Ease")

plt.show()
```



LINE CHART:

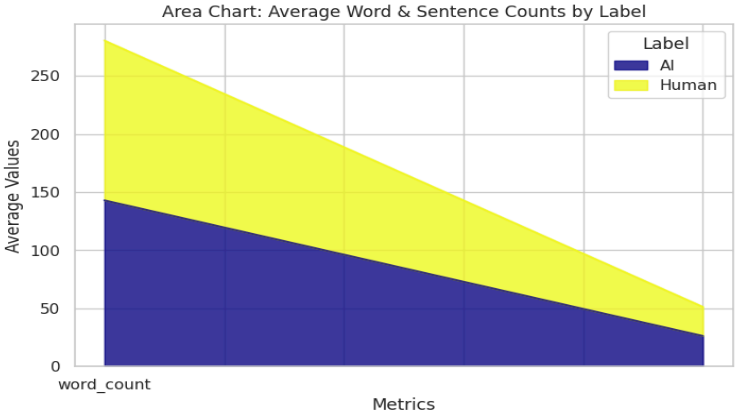
```
avg_readability = df.groupby('content_type')['flesch_reading_ease'].mean().reset_index()

plt.figure(figsize=(8, 5))

sns.lineplot(data=avg_readability, x='content_type', y='flesch_reading_ease', marker='o', color='teal')

plt.title("Average Readability by Content Type")

plt.xlabel("Content Type")

plt.ylabel("Average Flesch Reading Ease")

plt.xticks(rotation=45)
```



```
plt.show()
```

AREA CHART:

```
trend_df = df.groupby('label')[['word_count', 'sentence_count']].mean().T

trend_df.plot(kind='area', stacked=True, figsize=(8, 5), colormap='plasma', alpha=0.8)

plt.title("Area Chart: Average Word & Sentence Counts by Label")

plt.xlabel("Metrics")

plt.ylabel("Average Values")

plt.legend(title="Label")
```



```
plt.show()
```