Statistical Analysis of Human vs AI Generated Text

Project Group:
Subhadip Baidya (221092)
Souptik Majumder (221079)
Anant Srivastava (220133)

Contents

1 Overview

The project presents a detailed **statistical analysis** comparing **human-written** text with **AI-generated** text using entropy-based metrics. The main goal is to determine whether **entropy distributions** can help distinguish between human authorship and outputs from modern large language models.

2 Data Sources

2.1 Human-Generated Text

We used the *Human vs AI Text* dataset from Kaggle: https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text

2.2 Writing Prompts for AI Generation

Prompts for AI text generation were obtained from the *WritingPrompts* dataset available on HuggingFace: https://huggingface.co/datasets/euclaise/writingprompts

3 AI Models Used

The following large language models (LLMs) were used to generate AI-written text:

- **deepseek-r1:latest**
- **qwen2.5:7b**
- **llama3.2**

4 Methodology

The following steps were carried out:

1. Collected human-written text samples from Kaggle.
2. Sampled prompts from the WritingPrompts dataset.
3. Generated AI text using the selected LLMs.
4. Computed the following entropy-based metrics:
    - Token entropy
    - Normalized entropy
    - Sequence-average entropy
5. Visualized the entropy distributions using histograms.

All resulting plots are stored in the `figures/` folder.

5 Visualizations

5.1 Human-written Text Entropy Distribution

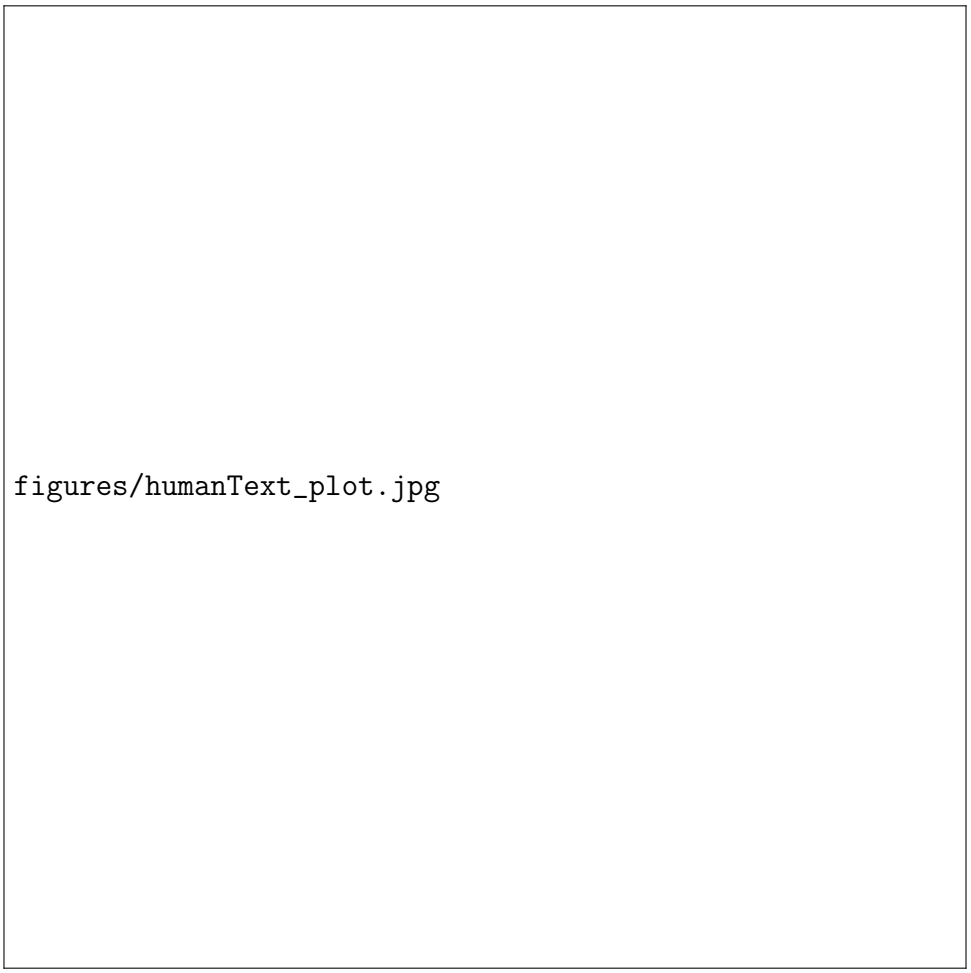

Figure 1: Entropy distribution for human-written text.

5.2 LLaMA 3.2 Text Entropy Distribution

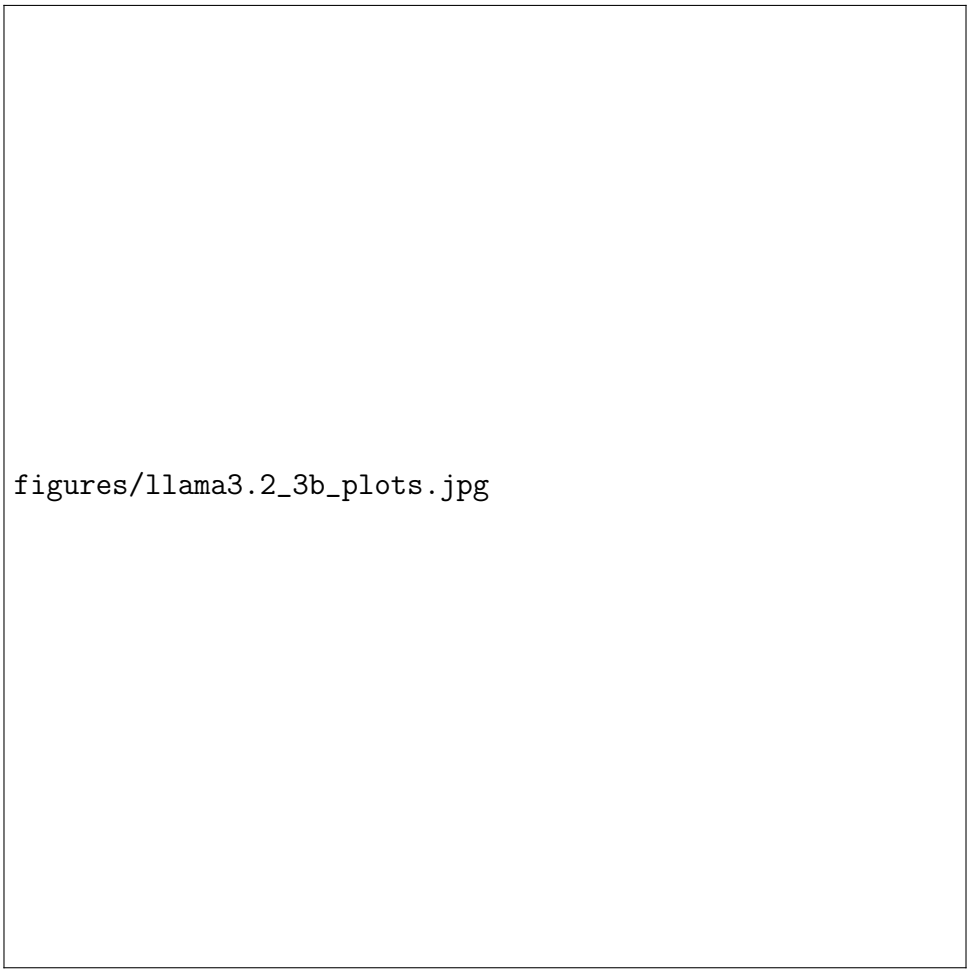

Figure 2: Entropy distribution for LLaMA 3.2 generated text.

5.3 Qwen2.5 7B Text Entropy Distribution

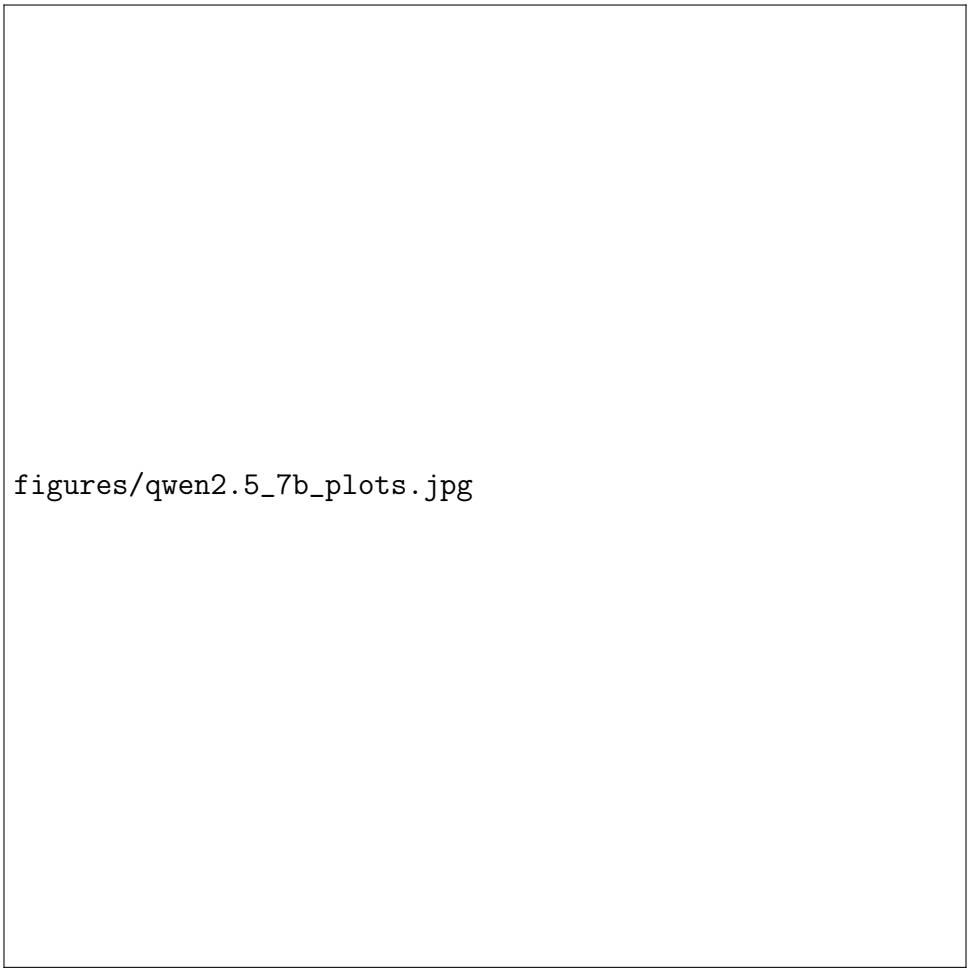

Figure 3: Entropy distribution for Qwen2.5 7B generated text.

6 AI vs Human Text Classification

We employed **Qwen-1.5B** as a feature extractor, following the layer selection strategy outlined in the paper *Text Fluoroscopy*. The extracted features were used to train three binary classifiers:

- Multi-Layer Perceptron (MLP)
- Random Forest (RF)
- Support Vector Machine (SVM)

The models were trained and evaluated on two test splits:

1. `gpt4-pub-gpt3`
2. `gpt4-writing-gpt3`

Metrics reported include **Accuracy, AUROC, Precision, Recall, F1-Score, and Confusion Matrix**.

7 Results

7.1 MLP Classifier Results

Validation

Metric	Score
Accuracy	0.8255
AUROC	0.8942
Precision	0.8702
Recall	0.7651
F1	0.8142
Confusion Matrix	[[132, 17], [35, 114]]

Test: gpt4-pub-gpt3

Metric	Score
Accuracy	0.5503
AUROC	0.5739
Precision	0.6271
Recall	0.2483
F1	0.3557
Confusion Matrix	[[127, 22], [112, 37]]

Test: gpt4-writing-gpt3

Metric	Score
Accuracy	0.7700
AUROC	0.8619
Precision	0.8403
Recall	0.6667
F1	0.7435
Confusion Matrix	[[131, 19], [50, 100]]

7.2 Random Forest Results

Validation

Metric	Score
Accuracy	0.7550
AUROC	0.8877
Precision	0.7000
Recall	0.8926
F1	0.7846
Confusion Matrix	[[92, 57], [16, 133]]

Test: gpt4-pub-gpt3

Metric	Score
Accuracy	0.4731
AUROC	0.4711
Precision	0.4737
Recall	0.4833
F1	0.4784
Confusion Matrix	[[69, 80], [77, 72]]

Test: gpt4-writing-gpt3

Metric	Score
Accuracy	0.7466
AUROC	0.8453
Precision	0.7434
Recall	0.7533
F1	0.7483
Confusion Matrix	[[111, 39], [37, 113]]

7.3 SVM Results

Validation

Metric	Score
Accuracy	0.8691
AUROC	0.9612
Precision	0.9167
Recall	0.8121
F1	0.8612
Confusion Matrix	[[138, 11], [28, 121]]

Test: gpt4-pub-gpt3

Metric	Score
Accuracy	0.5234
AUROC	0.4970
Precision	0.5479
Recall	0.2684
F1	0.3604
Confusion Matrix	[[116, 33], [109, 40]]

Test: gpt4-writing-gpt3

Metric	Score
Accuracy	0.8300
AUROC	0.8951
Precision	0.9159
Recall	0.7267
F1	0.8104
Confusion Matrix	[[140, 10], [41, 109]]

8 Problem Statement

The rapid rise of AI has led to a surge in machine-generated content, including both text and images, which are often used to spread misinformation and fake news. This creates a critical problem: distinguishing between human-written and AI-generated content.

The project aims to identify distinguishing characteristics between human and AI-generated text and determine whether it is possible to reliably separate them.

9 Steps to Achieve the Goal

Step 1: Statistical Analysis

Perform statistical analysis of text properties such as entropy and perplexity to examine differences between AI and human text.

Step 2: Deep Learning Approach

Based on the insights from the statistical study, implement a classifier combining raw text with statistical features.

The **Text Fluoroscopy** approach involves:

1. Using the model's vocabulary head $\phi(\cdot)$ to produce probability distributions q_0, q_N, q_j at different layers.
2. Computing KL divergence across layers to score their informativeness:
$$\text{Score}(j) = D_{\text{KL}}(q_N \mid\mid q_j)$$
3. Selecting the most informative layer $M = \arg\max_j \text{Score}(j)$.
4. Extracting the hidden state $h_{t-1}^{(M)}$ from layer M as the feature vector.
5. Feeding it into a binary classifier D trained via cross-entropy loss.

10 Datasets Used

- **Human Text:** Kaggle's Human vs AI Text dataset.
- **AI Text:** Generated using prompts from HuggingFace's WritingPrompts dataset with models: *deepseek-r1*, *qwen2.5-7b*, and *llama3.2*.

11 Evaluation Metrics

The performance of the classifiers was evaluated using:

- Accuracy
- Precision
- Recall
- F1-Score
- AUROC
- Confusion Matrix

12 Conclusion

The analysis demonstrates that entropy-based features and Qwen-1.5B layer representations can capture meaningful distinctions between human and AI-generated text. While performance varies across datasets and models, the SVM achieved the best validation results with **AUROC = 0.9612**, showing strong generalization potential.