Untitled

Du Shi, Kunning Zhang, Yixuan Yin

2026-02-05

Brief Introduction

Data preprocessing

The dataset contains around 3,500 complete English language texts presented in Stylometry form, grouped by author: Human, ChatGPT, Gemini and Llama. To solve this task, all large language model (LLM) texts were merged into a single category, resulting in a binary dataset of Human vs LLM. Then the data were randomly split into an 80% training set and a 20% testing set, and the texts were normalised to ensure comparability across texts of different lengths. Finally, the normalised 80% training set was used for exploratory data analysis.

Exploratory Data Analysis

	Number of texts	Number of Function words
Human	869	200
LLM	2608	200

The basic information of the training set is shown in Table 1. The training set contains 869 human written texts and 2,608 LLM written texts, with 200 function words for each text. Next, let's look at the mean difference in function words between LLM and Human texts.

Figure 1 shows the mean difference in function words, where the difference is LLM minus Human. We find that several function words have difference. In particular, V201 shows the largest mean difference, about 0.032, indicating that this function word is used very differently in human written and LLM written texts. This means V201 is an important feature for classifying human and LLM texts. Then, we used MDS visualisation to check the overall difference between Human and LLM texts.

As shown in Figure 2, the Human and LLM texts overlap but have different centroids, indicating partial separation and they may be separated in later analysis. Then, we used MDS Visualisation of Authors to explore the difference between human, ChatGPT, Gemini, and Llama texts.

In Figure 3, we find that LLM texts are on the right side and human texts are on the left side, indicating that human and LLM texts have different writing styles. In addition, different LLMs also have different writing styles. GPT is far away from Gemini and Llama. However, Gemini and Llama are close to each other, so they are more likely to be confused. After understanding the data, we start modelling.

Modelling

In this section, we will discuss two models 'K-nearest neighbors' and 'Discriminant Analysis' and compare to choose a more suitable model. Based on training data, we will use LOOCV to assess each performance. Particularly for KNN, a 10-fold cross validation will be used to choose a better value 'k'.

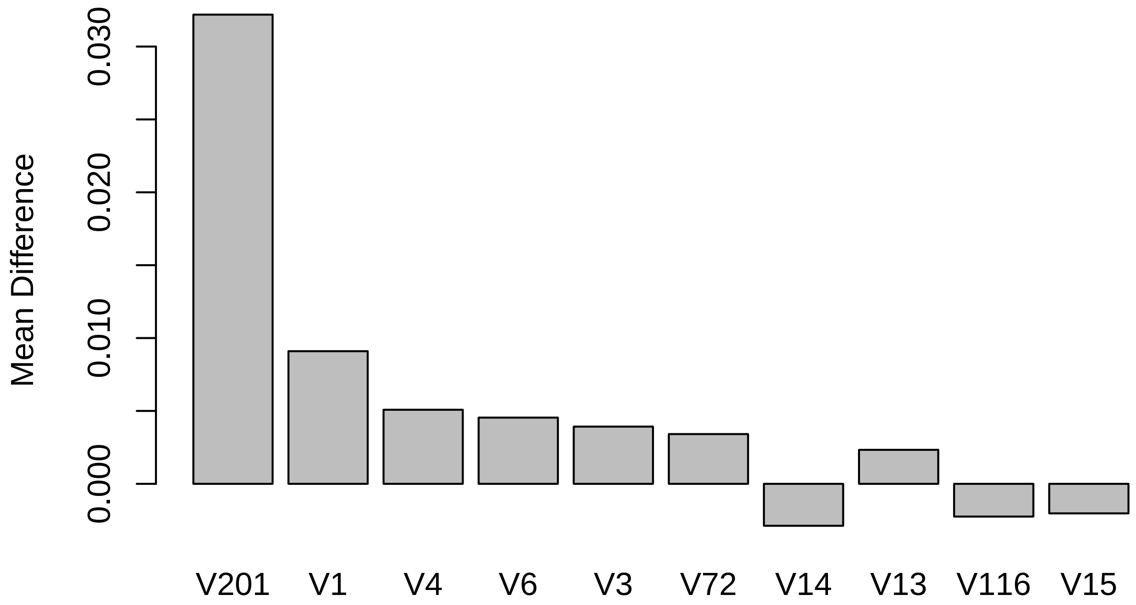


Figure 1: Top 10 Mean Difference in Function Words Between LLM and Human Texts

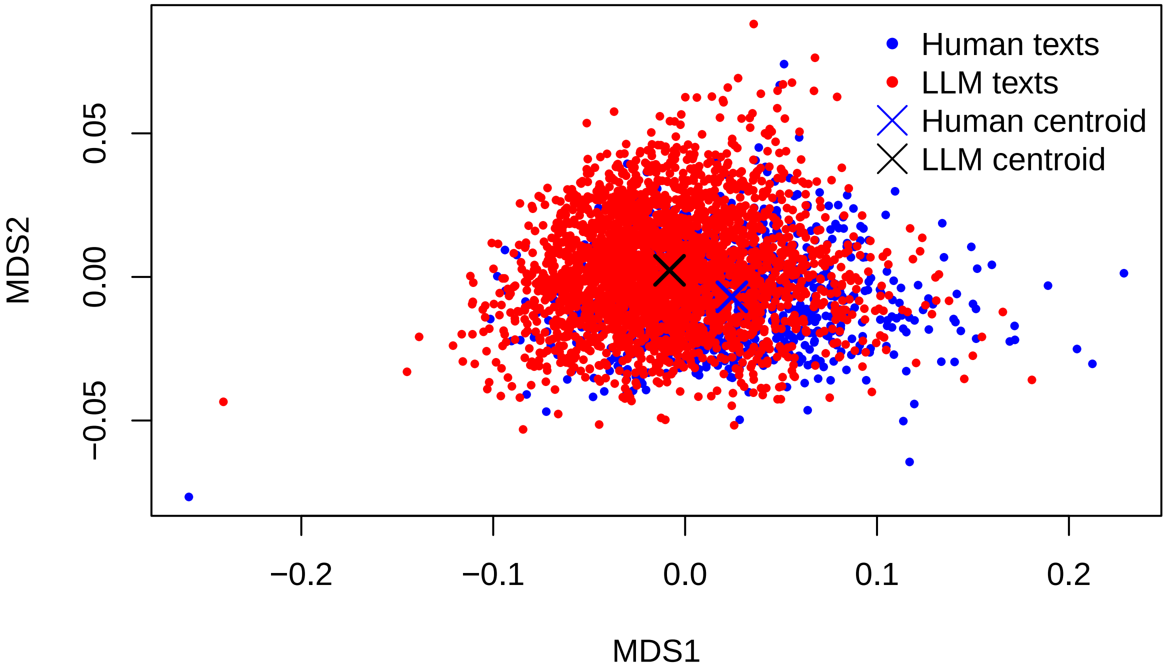


Figure 2: MDS Visualisation of Human and LLM Texts

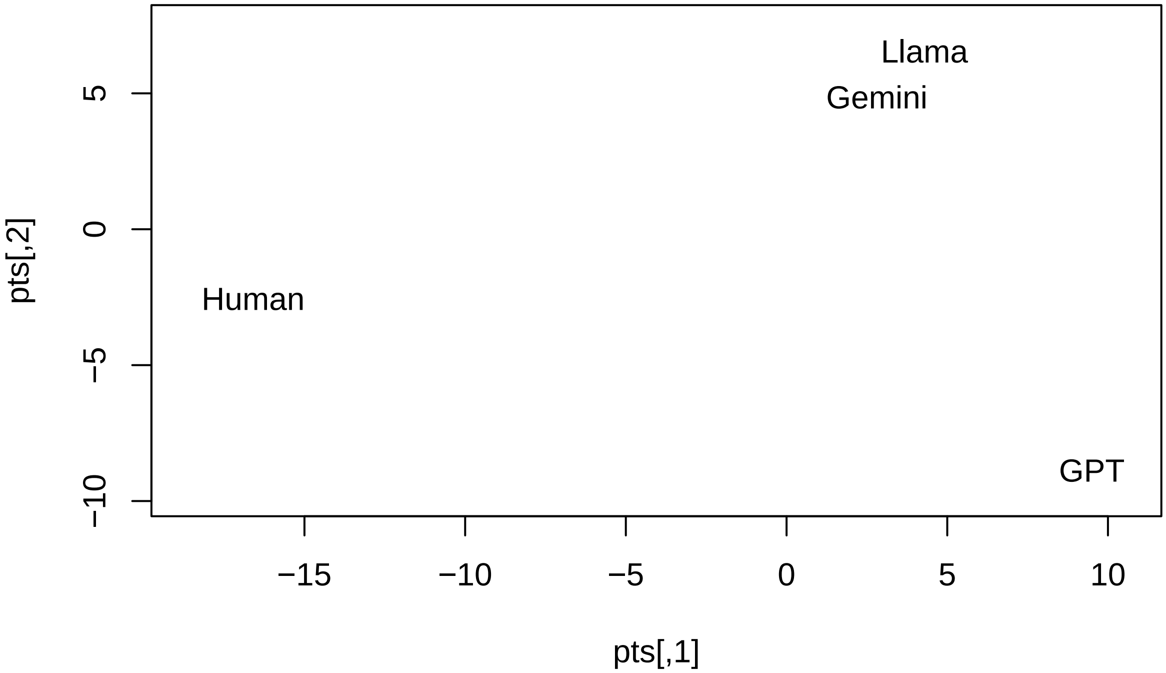


Figure 3: MDS Visualisation of Authors

KNN

'K-Nearest Neighbors' (KNN) assigns a class label to a test observation based on the majority label among its k closest training samples in feature space. Distance between observations is typically measured using Euclidean distance after feature standardization. Comparing to 'DA', 'KNN' makes no distributional assumptions and relies directly on local neighborhood structure.

However, its performance is sensitive to the choice of k, for example in case of k=1, an outlier of human data lays near AI's group can make test points of AI near it wrong labeled. Because of the number of AI training points is three times of Human, a very high k is not a good choice.

From figure 4 below, 'KNN' with k from 1 to 10 all gives similar accuracy around 0.8, but the human recalls are pretty low, lead to a low balanced accuracy. It means that many human data points are wrong classified to AI, which cause the fake high accuracy. We will discuss it later by confusion matrix in LOOCV after we choosing k=2 (k with highest human recall).

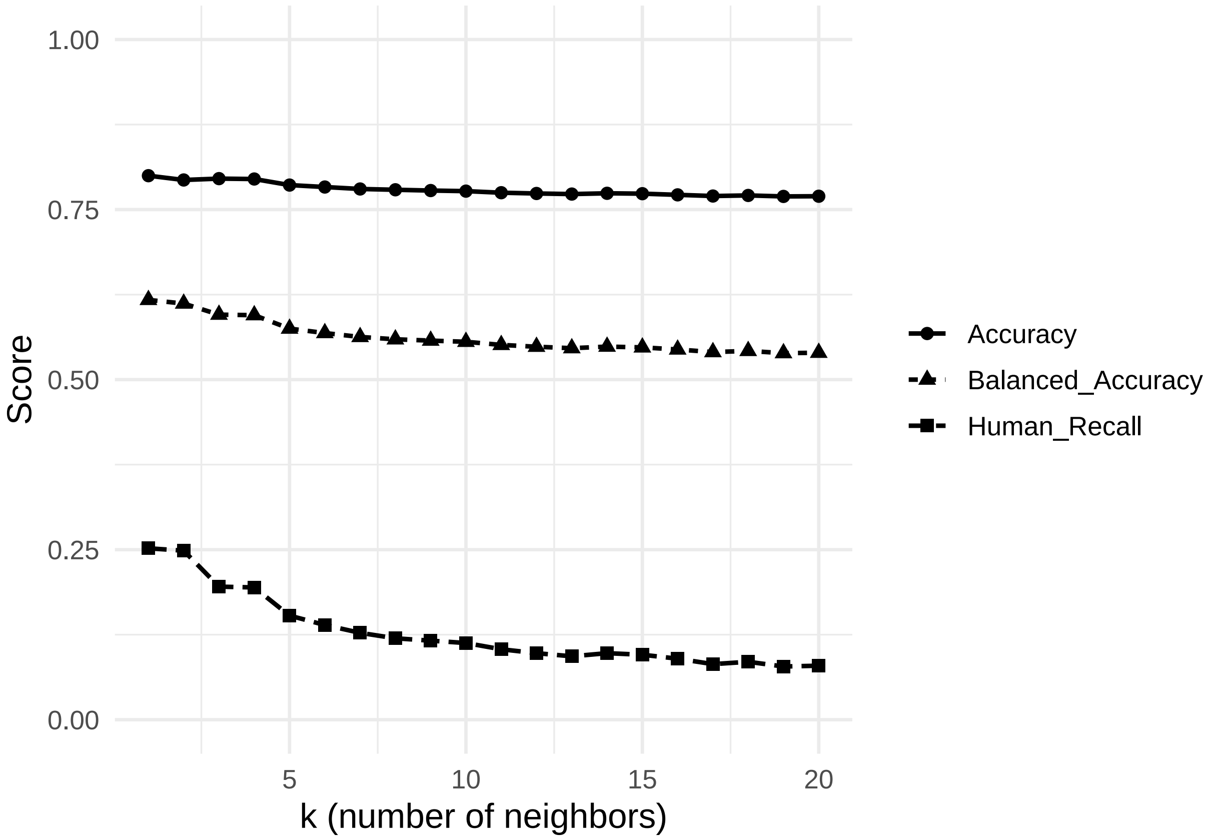


Figure 4: 10-fold cross-validated performance of KNN showing accuracy.

Table 2 below shows the LOOCV results of 'KNN' with k=2, 658 of human texts are classified to AI, which means that the model is bad on recognizing Human texts, most points are classified to AI.

Table 2: Confusion Matrix of LOOCV for KNN with k=2. ('1' is labeled as 'Human', '2' is labeled as 'AI')

1	2
220	65

1	2
649	2543

So the kappa value is pretty low so the true improvement of this model over chance is only moderate. And the sensitivity (Human recall) and balanced accuracy are pretty low.

The reason for this may come from the imbalanced number of data points together with high overlap of two classes.

Table 3: KNN(k=2) Performance — LOOCV

	Value
Accuracy	0.795
Kappa	0.294
Sensitivity	0.253
Specificity	0.975
Balanced Accuracy	0.614

DA

'Discriminant Analysis' assumes that observations from each class follow a multivariate normal distribution with means and variance from sample. Classification is performed by assigning each observation to the class with the higher probability. Because DA estimates a global decision boundary rather than relying on local neighbors, it is typically more stable KNN, especially when we have imbalanced number of data points.

Based on training data, Table 4 and 5 below show that 'DA' performs much better than 'KNN' in LOOCV. Only 131 from Human and 306 from AI are wrong classified. The accuracy is 0.874, with both high sensitivity and specificity. So 'DA' gives a fair classification with meaningful accuracy. Also kappa value (0.685) indicates substantial agreement.

Table 4: Confusion Matrix of LOOCV for DA. ('1' is labeled as 'Human', '2' is labeled as 'AI')

1	2
738	308
131	2300

Table 5: DA Performance — LOOCV

	Value
Accuracy	0.874
Kappa	0.685
Sensitivity	0.849
Specificity	0.882
Balanced Accuracy	0.866

Based on the validation results, we will choose discriminant analysis below to classify the testing data.

Testing Results

Now we start to classify testing data by discriminant analysis. The model is trained by trainset.

From the confusion matrix below, we can see that most texts are corrected labeled.

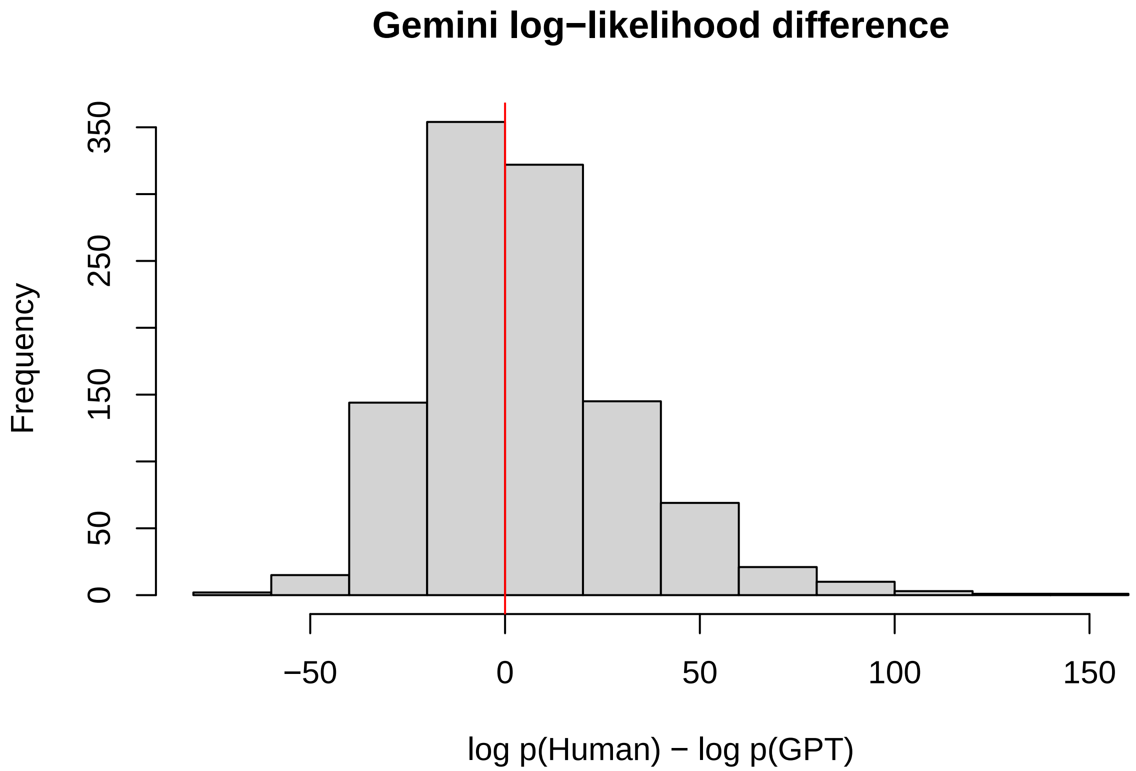
Table 6: Confusion Matrix of testing for DA. (‘1’ is labeled as ‘Human’, ‘2’ is labeled as ‘AI’)

1	2
177	76
41	577

Since testset is independent with trainset and separated randomly, we have evidence to say that for any given texts, the model is able to classify them between AI and Human at accuracy 95% confidence interval 0.841-0.888, with out obvious bias on AI or Human.

Table 7: DA Performance — Testing

	Value
Accuracy	0.866
Kappa	0.660
Sensitivity	0.812
Specificity	0.884
Balanced Accuracy	0.848
AccuracyLower	0.841
AccuracyUpper	0.888


Gemini log−likelihood difference
Frequency
0
50
150
250
350
−50
0
50
100
150
log p(Human) − log p(GPT)