Supervised Learning Model for LLM Text Detection

Du Shi, Kunning Zhang, Yixuan Yin

Executive summary

This paper will focus on classifying LLM texts and Human texts based on the difference on number of 200 function words they use in a text like 'a' and 'the', together with total other words. There will be 3 LLM models: GPT, Gemini and Llama, which will be all seen as LLM group.

We will talk about two model, 'K-nearest-neighbour' and 'Discriminant Analysis'. KNN labels points by their nearest group point. DA labels points by probability. We will introduce them later. Then we apply the chosen model DA on testing points to find a general results for future application and got an accuracy over 0.85.

In the end, as an extra research, we discuss the similarity between three LLM's and also comparing to Human. For example, we train the model by Human and GPT, and test on Gemini. The final results indicate that all three LLM texts differ significantly from human text, and these differences follow similar stylistic patterns. GPT has the most distinctive written style, Gemini and Llama are similar and both closer to human writing.

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 and model choosing.

Exploratory Data Analysis

Table 1: Basic Information of the Training Data

	Number of Texts	Function Words	Feature for other words
Human	869	200	1
LLM	2608	200	1

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 and one additional feature for all other words in each text. The mean differences in function words between LLM and Human texts is shown in figure 1

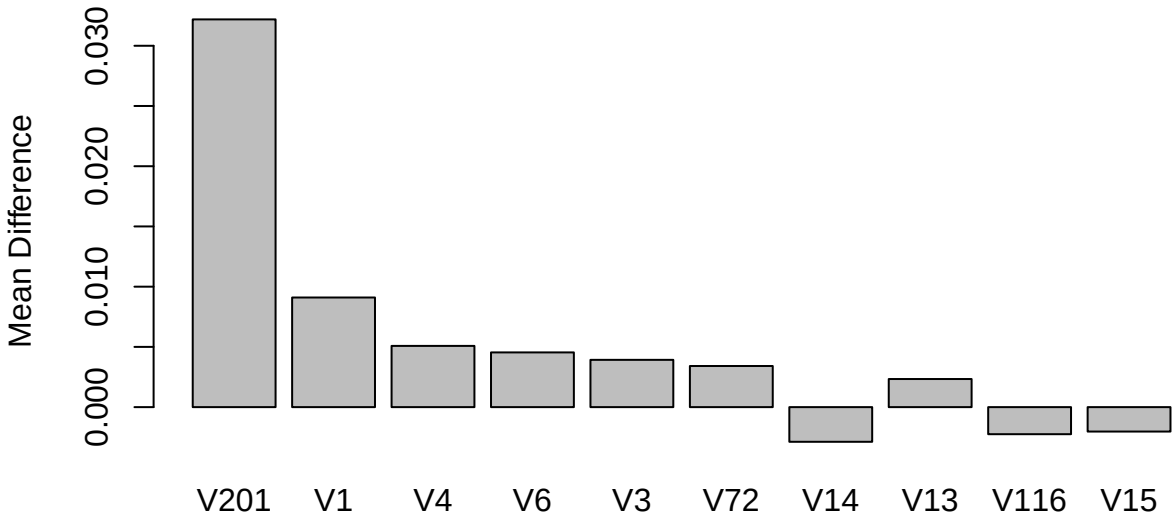

Figure 1: Top 10 Mean Differences in Function Words Between LLM and Human Texts

The feature V201 can be ignored, as its high value is primarily due to longer text length rather than meaningful mean difference. We find that several function words have difference. In particular, V1 shows the largest mean difference, about 0.01, indicating that this function word is used very differently in human written and LLM written texts. This means V1 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.

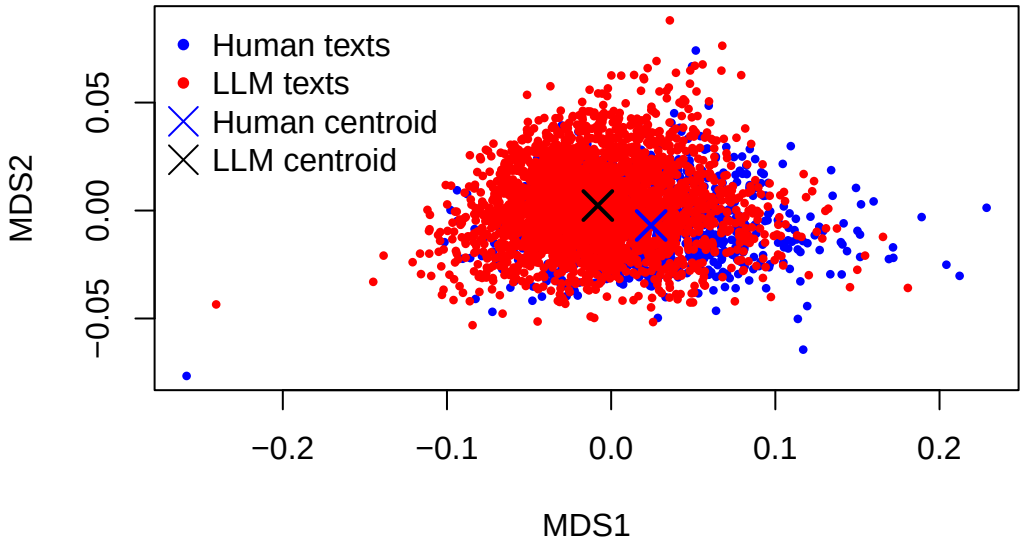

Figure 2: MDS Visualisation of Human and LLM Texts

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.

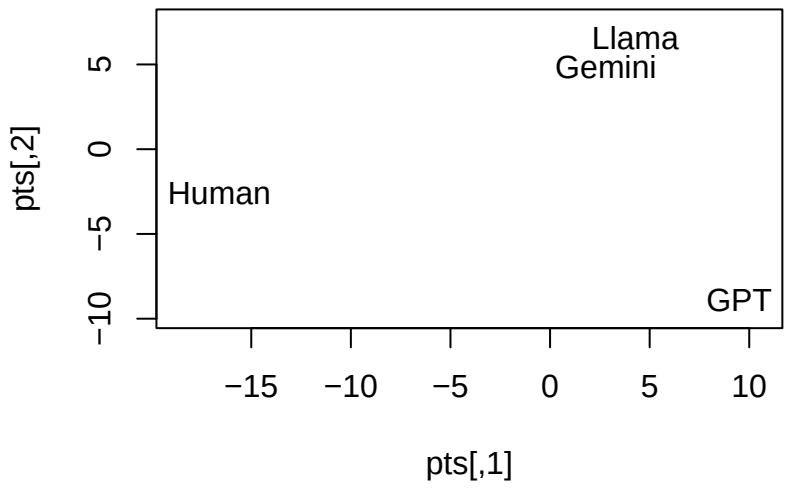

Figure 3: MDS Visualisation of Authors

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'.

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 LLM's group can make test points of LLM near it wrong labeled. Because of the number of LLM 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 LLM, 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).

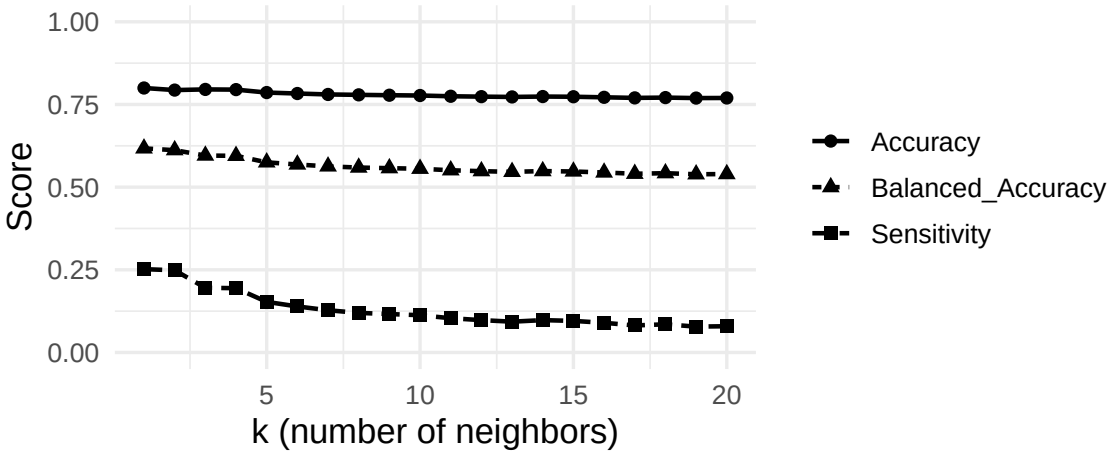

Figure 4: 10-fold cross-validated performance of KNN showing accuracy.

Table 2 below shows that the LOOCV results of 'KNN' with k=2, 644 of human texts are classified to LLM, which means that the model is bad on recognizing Human texts.

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

1	2
220	65
649	2543

Table 3 below shows that the kappa value is pretty low, so the true improvement of this model over chance is only moderate. And the sensitivity (Human recall) is much more lower than specificity (LLM recall), means that the model is biased on LLM.

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 308 from LLM 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 'LLM')

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 test set by discriminant analysis. The chosen 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 'LLM')

1	2
177	76
41	577

Since test set is independent with train set and separated randomly, we have evidence to say that for any given texts, the model is able to classify LLM text and Human text, achieving accuracy with a 95% confidence interval 0.841–0.888, without obvious bias on LLM 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

Extra Research

To see how the model performs if we use the "wrong" LLM to train the data, we further construct model by training on all Human written texts together with one of the three LLM texts, and testing on the other two LLM texts. (To make the results more obvious, we remove the human texts in the test set and only focus on whether the tested LLM can be classified as the trained LLM.)

Table 8: Accuracy If We Use The "Wrong" LLM To Train The Data

Train	Test	Accuracy
GPT vs Human	Gemini	0.474
GPT vs Human	Llama	0.613
Gemini vs Human	GPT	0.934
Gemini vs Human	Llama	0.902
Llama vs Human	GPT	0.925
Llama vs Human	Gemini	0.794

The accuracy in table 8 means the proportion of tested LLM generated texts as corresponding trained LLM, which varies significantly across different models. As shown in table 2, 0.474 and 0.613 indicates nearly half of the texts generated by Gemini and Llama are wrongly classified as human written, which shows a weak tendency of classifying texts. In contrast, the classifiers trained using Human and Gemini, as well as the one trained on Human and Llama, have obvious tendencies to classify the other two LLM-generated models.

The test results also demonstrate whether the LLM being tested has significant influence on classification outcomes. As shown in table 8, less than 10% of GPT-generated texts are classified as Human-written on both trained classifiers. Compared to GPT's performance, Llama and Gemini's results appeared a slightly weaker under the same classifiers.

Under this framework, The mean values of Gemini and Llama's features' normal distribution lay between GPT and Human and similar with each other, Gemini may tend to be near human more. GPT has the most distinctive written style. This result is exactly what figure 3 shows.