Model Summary

Competition Name: LLM - Detect AI Generated Text

Team Name: yellowleaf

Private Leaderboard Score: 0.892762

Private Leaderboard Place: 1597

GitHub repository: https://github.com/wcqy001028/LLM-Detect-AI-Generated-Text

Team Member

Name: wcqyfly

Location: Hangzhou, Zhejiang, China

Email: 2476226021@qq.com

Name: PhoenixKnight

Location: Tongling, Anhui, China

Email: GwendolynkAshe@outlook.com

Name: SilverLion33

Location: Hangzhou, Zhejiang, China

Email: TonyiWoodruffm@outlook.com

Name: X7Shadow

Location: Quanzhou, Fujian, China

Email: LeroylArcherx@outlook.com

1. Overview of the Approach

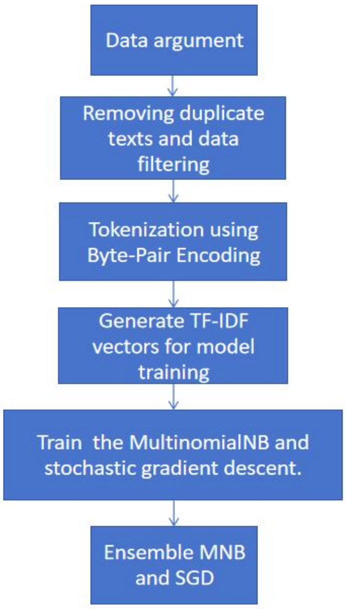
Data argument
Removing duplicate texts and data filtering
Tokenization using Byte-Pair Encoding
Generate TF-IDF vectors for model training
Train the MultinomialNB and stochastic gradient descent.
Ensemble MNB and SGD

Fig. 1 The pipeline of our solution

Our solution includes **six parts** based on public work.

(https://www.kaggle.com/code/batprem/llm-daigt-excluded-prompts?scriptVersionId=158926419)

(1).Data argument.

The mainstream approach in competition is to seek or generate diverse datasets of student writing and large language model-generated data for data augmentation. Many Kagglers have generously shared their own datasets. Our approach also involves searching for external open-source data, and we have utilized the following external datasets:

https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset

https://www.kaggle.com/datasets/alejopaullier/argugpt

https://www.kaggle.com/datasets/kagglemini/train-00000-of-00001-f9daec1515e5c4b9

https://www.kaggle.com/datasets/pbwic036/commonlit-data

https://www.kaggle.com/datasets/wcqyfly/argu-train

(2).Removing duplicate text and data filtering.

This part is the same as the public work.

(https://www.kaggle.com/code/batprem/llm-daigt-excluded-prompts?scriptVersionId=158926419)

(3).Tokenization using Byte-pair Encoding.

This part is the same as the public work.

(https://www.kaggle.com/code/batprem/llm-daigt-excluded-prompts?scriptVersionId=158926419)

(4).Generate TF-IDF vectors for model training.

We adjusting parameters such as min_df, max_df, and max_features to effectively reduce the feature dimensionality and minimize the impact of noisy data.

(5).Train the MultinomialNB and stochastic gradient descent.

We adjusted the parameter alpha which means the additive smoothing for MultinomialNB and the train steps for SGD.

(6).Ensemble MNB and SGD.

We have attempted to combine Multinomial Naive Bayes (MNB) and Stochastic Gradient Descent (SGD) using different weights.

2.Run time optimization operation

To combine the dataset in other notebook (https://www.kaggle.com/wcqyfly/notebook95c85fa3c6).

Reduced feature size as mentioned above. We adjusting parameters such as min_df, max_df, and max_features.

Remove time-consuming models. We found that after using lightgbm and catboost, not only does it take more time, but the LB score also decreases. We decided to remove these two models. However, the fact proves that this choice is wrong, which makes the PB score very low. But fortunately, we have the opportunity to obtain Efficiency Price.

3.Details of the submission

Version	Public Score	Private Score
1	0.962507	0.861747
2	0.962465	0.853448
3	0.967997	0.891007
4	0.967267	0.898650
5	0.967885	0.890983

4.Which did not work

We adjusted the N_grams for TF-IDF, it did not work on both public score and private source.

5.Our code

Our code can be access by kaggle or github. The links are as follows:

1.kaggle:

https://www.kaggle.com/code/wcqyfly/fork-of-fork-of-fork-of-llm-daigt-analyse-e-db6333

note: It should be noted that because the number of test sets is less than 3, running all directly will cause the code to report an error, but after submitting, when the test set is replaced with a hidden test set, the code will be run correctly and get the result.

2.github: https://github.com/wcqy001028/LLM-Detect-AI-Generated-Text

note: If the number of data in the test_essays.csv is less than 5, the min_df is set to 1 and the model is not trained which only used for debugging. Conversely, when the number of data in test_essays.csv is greater than 5, the min_df is set to 2 and the model will be trained and will generate prediction results. <br/>

6. Reference

1.https://www.kaggle.com/code/batprem/llm-daigt-excluded-prompts?scriptVersionId=158926419

2.https://www.kaggle.com/datasets/thedrcat/daigt-v2-train-dataset

3.https://www.kaggle.com/datasets/alejopaullier/argugpt

4.https://www.kaggle.com/datasets/kagglemini/train-00000-of-00001-f9daec1515e5c4b9

5.https://www.kaggle.com/datasets/pbwic036/commonlit-data

6.https://www.kaggle.com/datasets/wcqyfly/argu-train

7.https://www.kaggle.com/competitions/llm-detect-ai-generated-text/discussion/468908

8.https://www.kaggle.com/competitions/llm-detect-ai-generated-text/discussion/455711