Natural Language Processing (CS310) Final Project Report

Detecting Human and LLMs-generated Texts using Supervised Learning-based and Zero-shot Approach across Diverse Domains

Fitria Zusni Farida, Sreyny Tha, Tan Hao Yang

Taught by Prof. Yang Xu

Department of Computer Science and Technology

Southern University of Science and Technology

{12112351, 12113053, 12212027}@mail.sustech.edu.cn

1. Introduction

To effectively detect large language models (LLMs)-generated texts, especially to distinguish them from real human-written ones, is becoming a more and more important task. The task can be approached with two technical paths: 1) **supervised learning**-based detection; 2) likelihood metrics-based **zero-shot** detection. The former is similar to building a text classification model for tasks such as sentiment analysis etc., which can be done by fine-tuning a transformer encoder-based model (e.g., BERT) on an annotated dataset with binary labels (e.g., "0" for human-written and "1" for LLM-generated). The main advantage of this approach is that a supervised learning model can perform well provided with sufficient amount of data, and will be useful for a focused task-domain (e.g., news, fictions etc.). The limitation is also obvious – it is not a generic method, which means a detection model trained on one type of text data may fail on others, that is, relatively poor out-of-domain (OOD) performance. The latter approach, likelihood-based zero-shot detection, is a more generic solution – the detection algorithm/pipeline developed for one text-domain/languages/LLM can also work well on others, that is, better overall OOD performance. (This introduction is referenced from the project docs)

The goal of this project is two-fold:

(1) Implement a series of supervised learning-based detection models, and test their performances under the OOD condition.
(2) Implement zero-shot detection methods, and test it on the same setting.

Our documentation is in this github [1]repository:

[1] https://github.com/always-hy/CS310_NLP_Project

2. Experimental Setups

A. English text classification

1) Datasets and data preprocessing

The dataset used is the provided *ghostbuster* dataset. Data preprocessing includes tidying up the given .txt files into json files and dropping samples which are empty. Finally, the datasets used are organized as follows:

- **gb-dataset:** The dataset used for training contains 21,994 samples, specifically from three domains: 8,000 samples of *reuter*, 7000 samples of *wp*, and 6,994 samples of *essay*. The dataset contains texts and labels with 2,994 samples of human written text and 19,000 samples of AI-generated (gpt, gpt_prompt1, gpt_prompt2, gpt_semantic, gpt_semantic, and claude). The dataset distribution is as follows.

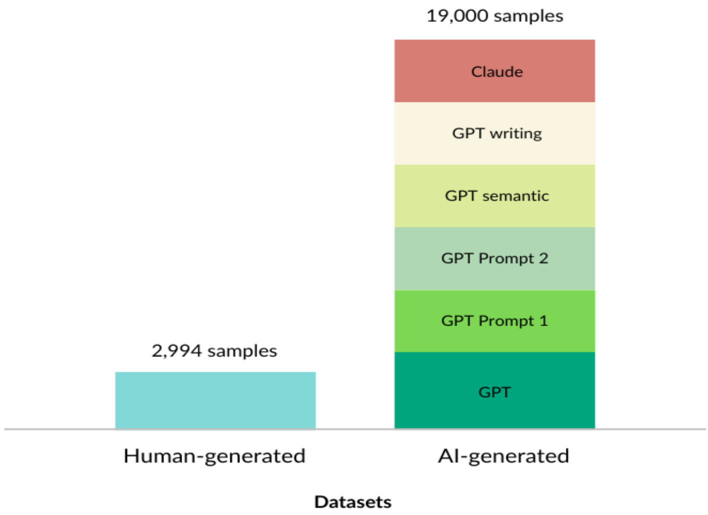

Figure 1. gb-dataset distribution for supervised fine-tuning from human and AI-generated texts in *reuter, wp, and essay* domain

- **gb-pair-dataset:** The dataset contains human and AI-generated (gpt, gpt_prompt1, gpt_prompt2, gpt_semantic, gpt_semantic, and claude) pairs with respect to its context. The domains are the same with gb-dataset (200 pair texts from *essay*, 1,000 pair texts from *reuter*, and 200 pair texts from *wp*).
- **gb-ood-dataset:** out-of-domain datasets are used to test the model's performance out of the domain used during the training. The domains are specifically from: *poet* (1,800 samples) and *mental health counseling conversation* (602 samples).

- **gb-pair-comp-dataset:** This dataset is used for unsupervised-based zero-shot to compare with supervised-based. It contains human and AI-generated (using GPT-4o-mini) with respect to its context. The domains are specifically from: *poet* (302 pairs) and *mental health counseling conversation* (200 pairs).

2) Methodology

(a) Supervised learning-based classification

For this task to classify human-generated vs AI-generated texts implementing supervised-based classifiers, we will fine-tune the **BERT**[2] (specifically bert-base-uncase variant) model in **gb-dataset.** To test the performance out of the trained domain, **gb-ood-dataset** will be used.

(b) Zero-shot classification

For this task to classify human-generated vs AI-generated texts implementing unsupervised-based classifiers, we will implement zero-shot approach **FourierGPT**[3] **pairwise heuristic-based** with **GPT-2 model** to obtain the NLL score (negative likelihood) and applying zscore normalization. The dataset used for this implementation is **gb-pair-dataset**. Also, another testing will be performed using **gb-pair-comp-dataset**.

B. Chinese text classification

1) Datasets and data preprocessing

The dataset used is the provided *face2* dataset. Data preprocessing includes tidying up the given .txt files into json or csv files and dropping samples which are empty. Finally, the datasets used are organized as follows:

- **face-dataset**: The dataset used for training is specifically from three domains: *news*, *webnovel*, and *wiki*. It contains normally distributed texts and labels with 14,967 samples of AI-generated (model to generate is not specified) chinese text and 13,145 samples of human-generated chinese text.

---

[2] https://huggingface.co/google-bert/bert-base-uncased

[3] https://github.com/CLCS-SUSTech/FourierGPT

- **face-pair-dataset:** Pairwise (human and AI-generated pairwise with respect to the input) used is specifically from three domains (same as above). It contains 4,530 pairwise chinese texts from *news* domain, 4997 pairwise chinese texts from *webnovel* domain, and 3,607 pairwise chinese texts from *wiki* domain.
- **face-ood-dataset:** Out-of-domain datasets are used to test the model's performance out of the domain used during the training. The domains are specifically from: *finance* (3,543 samples)**,** *medicine* (1,195 samples)**,** and *law* (2,148 samples)**.**
- **face-pair-comp-dataset:** This dataset is used for unsupervised-based zero-shot to compare with supervised-based. It contains human and AI-generated pairwise. The domains are specifically from: *finance* (1,384 pairs)*, medicine* (1,074 pairs)*,* and *law* (425 pairs).

2) Methodology

(a) Supervised learning-based classification

For this task to classify human-generated vs AI-generated texts implementing supervised-based classifiers, we will fine-tune the **BERT**[4] (specifically bert-base-chinese variant) model in **face-dataset.** To test the performance out of the trained domain, **face-ood-dataset** will be used.

(b) Zero-shot classification

For this task to classify human-generated vs AI-generated texts implementing unsupervised-based classifiers, we will implement zero-shot approach **FourierGPT pairwise heuristic-based** with **GPT-2**[5] **model** to obtain the NLL score (negative likelihood) and applying zscore normalization. The dataset used for this implementation is **face-pair-dataset** and **face-pair-comp-dataset**.

3. Experiment Results

A. English text classification

(a) Supervised-based classification

[4] https://huggingface.co/google-bert/bert-base-chinese
[5] https://huggingface.co/docs/transformers/model_doc/gpt2

Fine-tuning BERT model (bert-base-uncased) using following hyperparameters:

```
test_size = 0.2
num_epochs = 10
learning_rate = 2e-5
max_length = 512
batch_size = 64
```

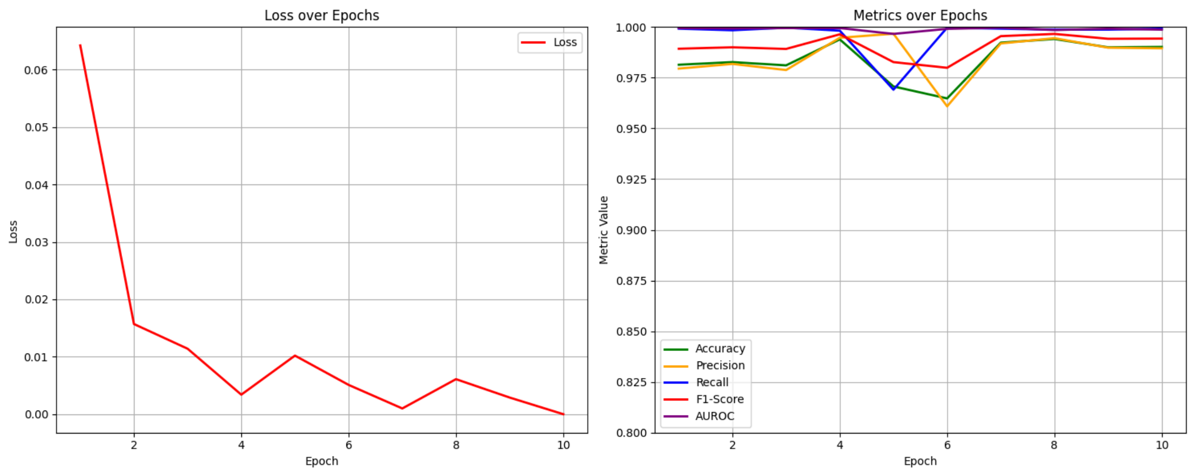

Figure 2. Training loss fine tuning BERT in **gb-dataset** and accuracy, precision, recall, F1-score, and AUC-ROC in validation set for English texts classification

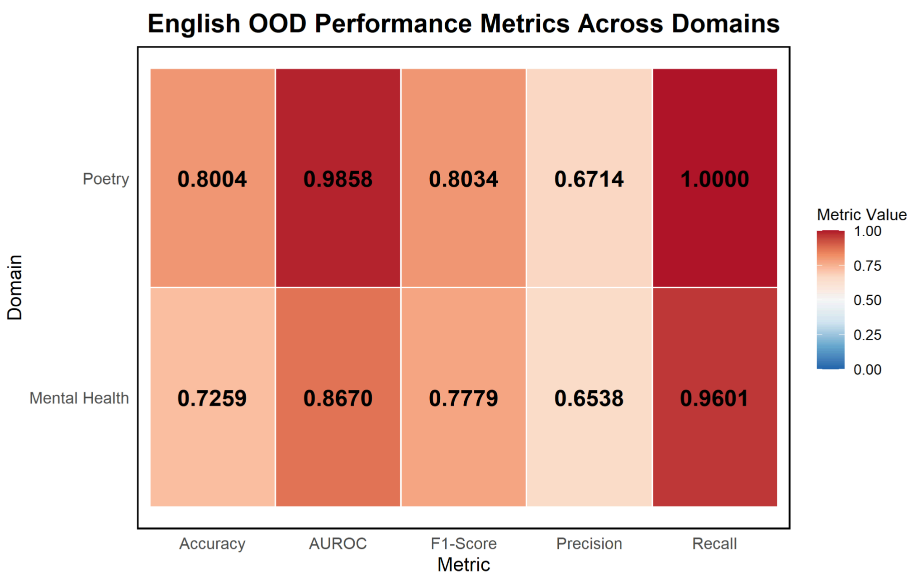

Figure 3. Accuracy, precision, recall, F1-score, and AUC-ROC out-of-domain in **gb-ood-dataset** for English texts classification

(b) Zero-shot classification

Zero-shot Accuracy (%) English text

	Essay	Reuter	Wp
GPT	k = 5 | 62.50%	k = 5 | 65.20%	k = 24 | **79.00%**
GPT prompt 1	k = 10 | 62.50%	k = 48 | 59.80%	k = 24 | **73.50%**
GPT prompt 2	k = 6 | 56.00%	k = 44 | **71.00%**	k = 9 | 69.00%
GPT semantic	k = 5 | 58.50%	k = 46 | 66.10%	k = 30 | **75.00%**
GPT writing	k = 5 | 62.00%	k = 11 | 57.10%	k = 21 | 62.00%
Claude	k = 4 | 64.00%	k = 4 | 67.10%	k = 9 | 65.60%

Best k | accuracy   Higher = model   Higher = human

Figure 4. Accuracy and best k implementing FourierGPT in **gb-pair-dataset** across various LLMs and domains for English texts classification.
Note: best k is the number of the first k frequency components (from low to high) that produces the best accuracy

Zero-shot Accuracy (%) English text

Dataset	besk k | accuracy
Mental health conversation	k = 18 | **73.09%**
Poem	k = 50 | **87.27%**

Higher = model   Higher = human

Figure 5. Accuracy and best k implementing FourierGPT in **gb-pair-comp-dataset** for English texts classification.
Note: best k is the number of the first k frequency components (from low to high) that produces the best accuracy

B. Chinese text classification

(a) Supervised-based classification

Fine-tuning BERT model (bert-base-chinese) using following hyperparameters:

```
test_size = 0.2
num_epochs = 10
learning_rate = 2e-5
max_length = 512
batch_size = 64
```

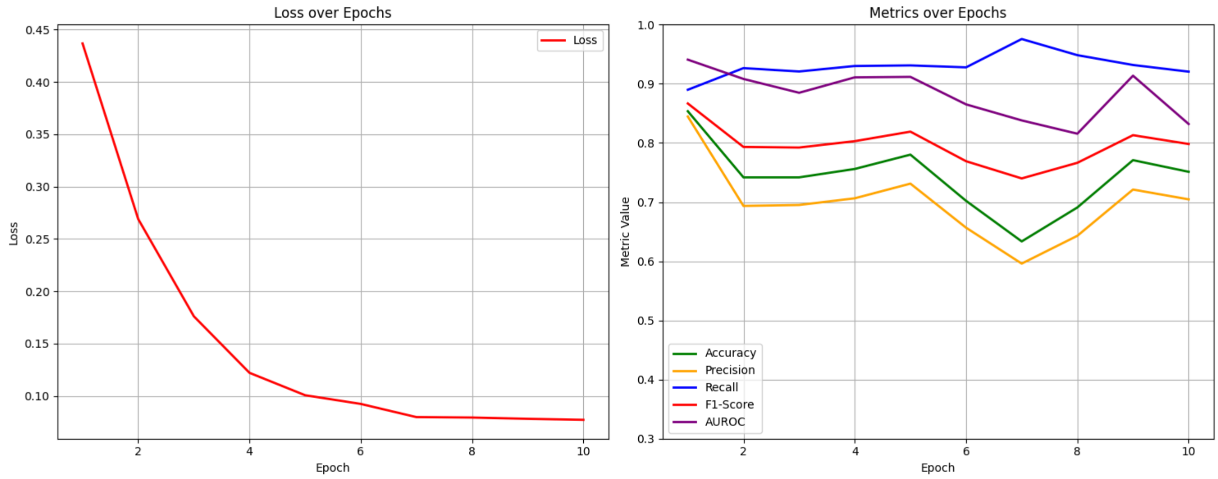

Figure 6. Training loss fine tuning BERT in **face-dataset** and accuracy, precision, recall, F1-score, and AUC-ROC in validation set for Chinese texts classification

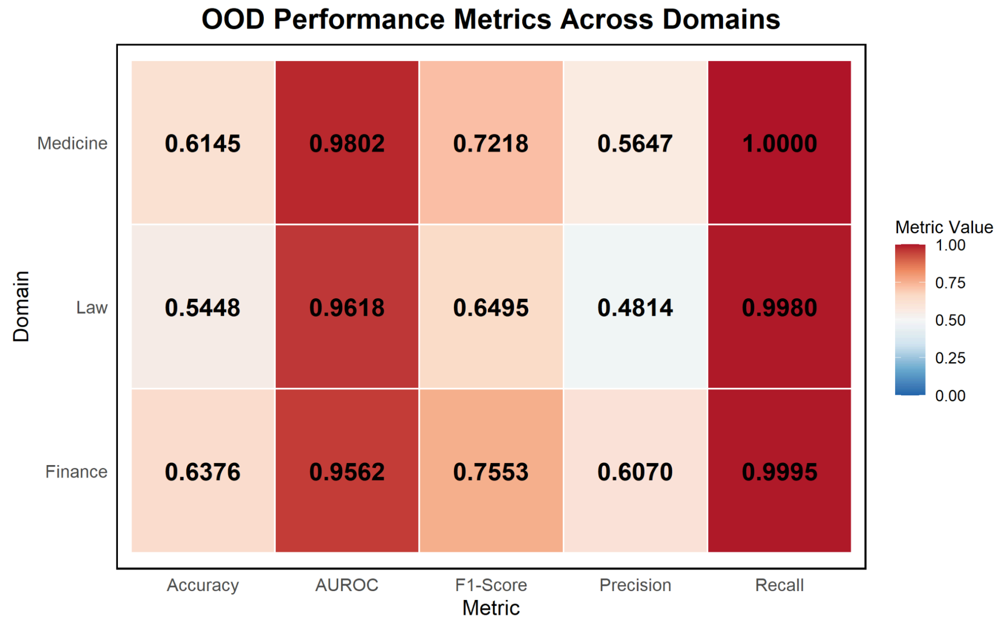

Figure 7. Accuracy, precision, recall, F1-score, and AUC-ROC out-of-domain in **face-ood-dataset** for English texts classification

(b) Zero-shot classification

Zero-shot Accuracy (%) Chinese text

Dataset	besk k | accuracy
News	k = 4 | **60.32%**
Webnovel	k = 1 | **60.83%**
Wiki	k = 5 | 59.06%

Higher = model Higher = human

Figure 8. Accuracy and best k implementing FourierGPT in **face-pair-dataset** for Chinese texts classification.
Note: best k is the number of the first k frequency components (from low to high) that produces the best accuracy

Zero-shot Accuracy (%) Chinese text

Dataset	besk k | accuracy
Finance	k = 8 | **85.45%**
Law	k = 32 | **92.71%**
Medicine	k = 29 | 62.31%

Higher = model Higher = human

Figure 9. Accuracy and best k implementing FourierGPT in **face-pair-comp-dataset** for Chinese texts classification.
Note: best k is the number of the first k frequency components (from low to high) that produces the best accuracy

4. Conclusion

1) Supervised fine-tuning BERT performed well, especially for in-domain datasets for both English and Chinese classification tasks.
2) Supervised Fine-tuning BERT performed poorly in OOD compared with in-domain dataset.
3) Zero-shot FourierGPT approach performed worse compared to supervised fine-tune for in-domain dataset.
4) Zero-shot FourierGPT was better compared with fine-tuning BERT in OOD.
5) In zero-shot FourierGPT, the result showed model-generated tends to have higher average power value on the selected frequency components.