# Detection of AI-generated texts

Detect Detection of AI-generated texts for SuperAI3 Hackathon

301503 - สุพศิน

300190-กัญจน์

# preprocess

- ใช้แค่ข้อมูลในคอลัมน์ของ Summary
- label Summary ของ Human เป็น 0 ของ AI เป็น 1
- concat and train_test_split(shuffle=True)

| | summary | label |
|---|---|---|
| 0 | We show that rare but catastrophic failures ma... | 0 |
| 1 | In this paper we propose a hierarchical archit... | 0 |
| 2 | We propose Episodic Backward Update, a novel d... | 0 |
| 3 | A new RL algorithm called Interior Policy Diff... | 0 |
| 4 | We investigate the modularity of deep generati... | 0 |

```
df_test.head()
```

| | id | abstract | summary1 | summary2 |
|---|---|---|---|---|
| 0 | test_0001 | This paper addresses the problem of evaluating... | The field of few-shot learning has recently se... | We show that rare but catastrophic failures ma... |
| 1 | test_0002 | We explore efficient neural architecture searc... | Convolutional neural networks (CNNs) have been... | In this paper we propose a hierarchical archit... |
| 2 | test_0003 | We propose Episodic Backward Update - a new al... | We propose Episodic Backward Update, a novel d... | One of the distinguishing aspects of human lan... |
| 3 | test_0004 | Animals develop novel skills not only through ... | A new RL algorithm called Interior Policy Diff... | Distributed computing can significantly reduce... |
| 4 | test_0005 | Deep generative models such as Generative Adve... | We investigate the modularity of deep generati... | We investigate methods for semi-supervised lea... |

**OpenAI Report**

November, 2019

---

# Release Strategies and the Social Impacts of Language Models

---

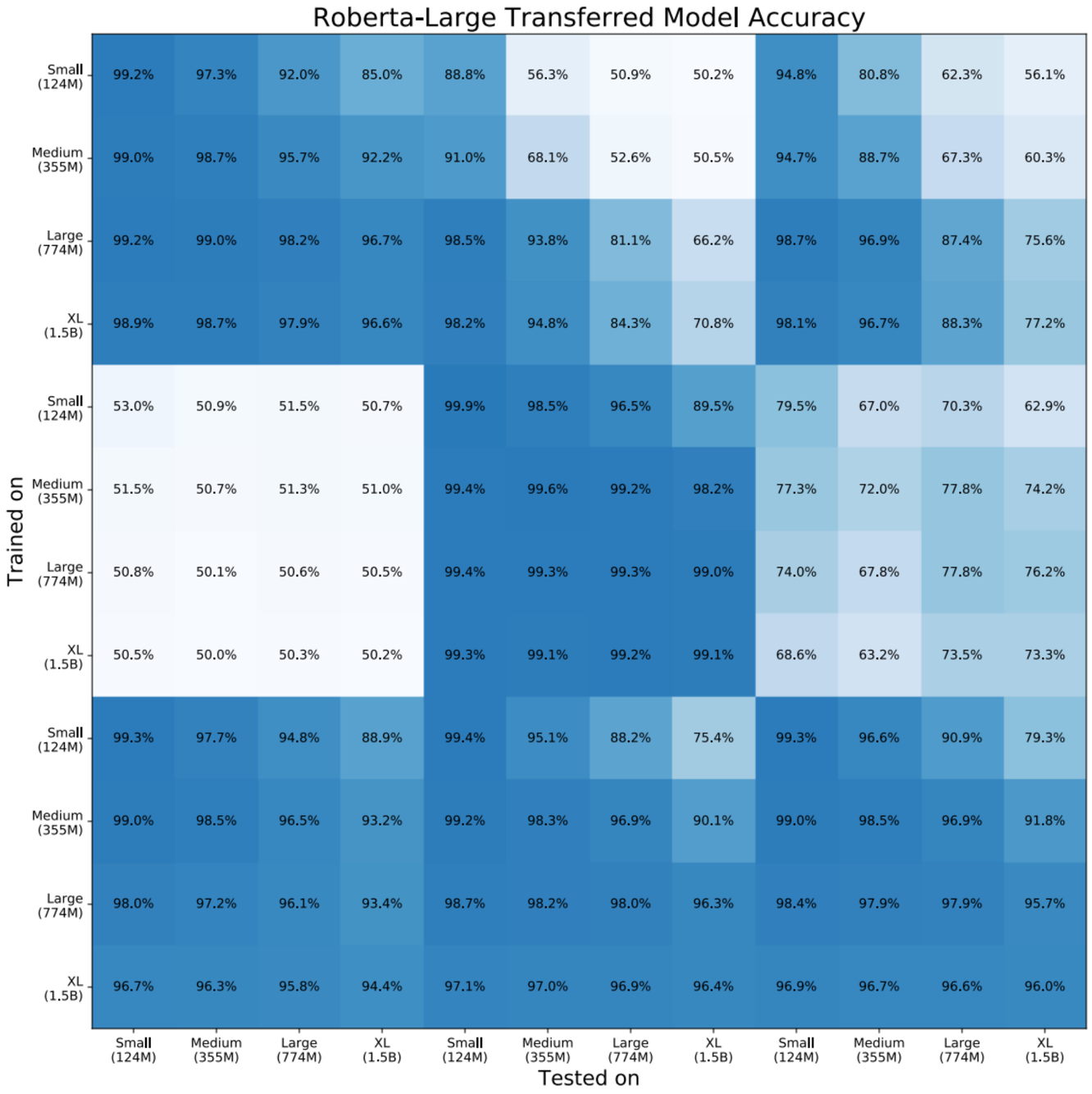
**1. Simple classifiers: Uses classifiers trained from scratch.**

**2. Zero-shot detection: Uses a pre-trained generative model (e.g., GPT-2 or GROVER) to outputs from itself or similar models**

**3. Fine-tuning based detection:**

logistic regression detector on TF-IDF(term frequency-inverse document frequency)
= 74-88% acc

83-85% acc

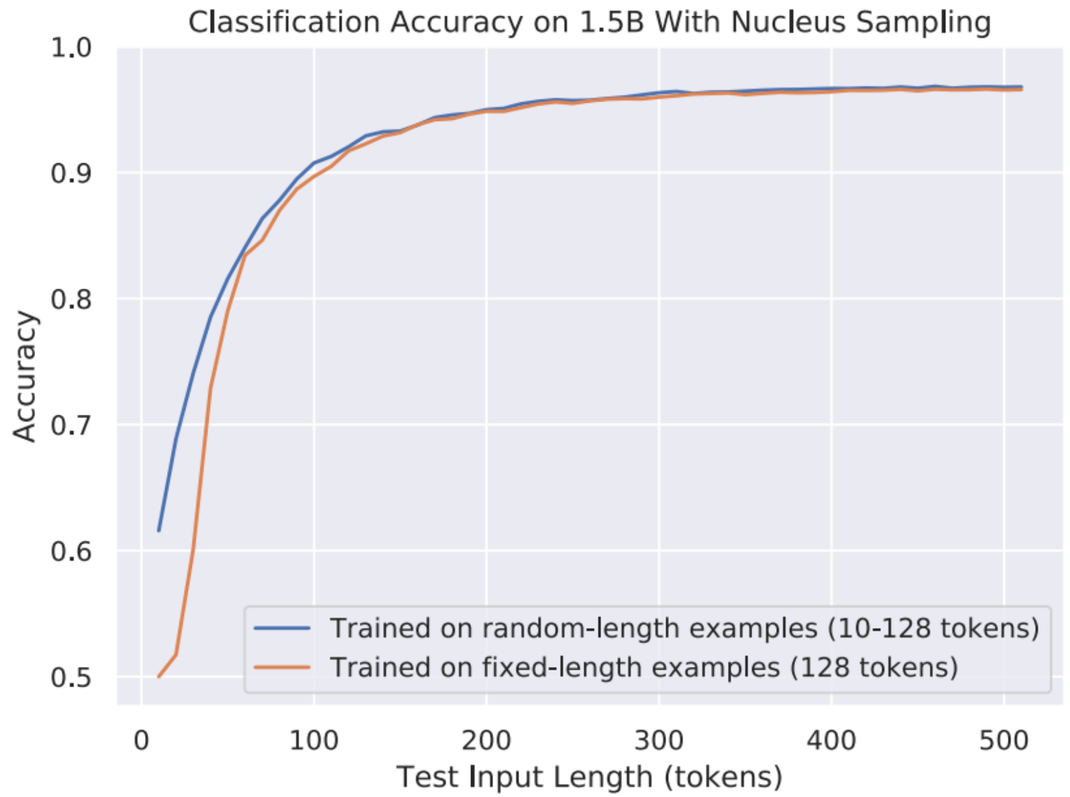



To develop a robust detector model they performed an analysis of the model's transfer performance. The 12-by-12 matrix shows the transfer accuracy with respect to the combination of four model sizes (124M, 355M, 774M, and 1.5B) and three sampling methods (Temperature = 1, Top-K = 40, and nucleus sampling with the Top-P sampled uniformly between 0.8 and 1.0).

- GPT-XL (1.5B parameters)
- GPT-Large (774M parameters
- GPT-Medium (355M parameters)
- GPT-2 (124M parameters)

imply that larger models' outputs will become more difficult to detect

The detection accuracy becomes higher for longer text,
roughly surpassing 90% accuracy
at 100 RoBERTa tokens



**roberta-base-openai-detector** ♡ like 71

Text Classification | PyTorch | TensorFlow | JAX

arxiv:1904.09751 | arxiv:1910.09700 | License: mit

```
model_args = {
    "reprocess_input_data": True,
    'train_batch_size':48,
    "use_early_stopping": True,
    "early_stopping_delta": 0.01,
    "early_stopping_metric": "mcc",
    "early_stopping_metric_minimize": False,
    "early_stopping_patience": 5,
    "evaluate_during_training_steps": 500,
    "fp16": False,
    "overwrite_output_dir":True,
    'use_cached_eval_features' : False,
    'max_seq_length': 128,
    'no_cache': True,
    "num_train_epochs": 10
}

model = ClassificationModel(
     "roberta",
     "roberta-base-openai-detector",
     use_cuda=torch.cuda.is_available(),
     num_labels=2,
)
```

```
0%  8/3758 [00:02<14:58, 4.17it/s]
Epoch 1 of 1: 100%  1/1 [01:04<00:00, 64.31s/it]
Epochs 0/1. Running Loss: 0.0000: 100%  470/470 [00:58<00:00,
```

| | id | sum1 | sum2 |
|---|---|---|---|
| 0 | test_0001 | 1 | 0 |
| 1 | test_0002 | 1 | 0 |
| 2 | test_0003 | 0 | 1 |
| 3 | test_0004 | 0 | 1 |
| 4 | test_0005 | 0 | 1 |

# postprocess

- แยก Predict summary1 และ summary2
- แปลงให้อยู่ในรูปของ

class 0 : summary1 = human, summary2 = AI (0,1)
class 1 : summary1 = AI, summary2 = human (1,0)

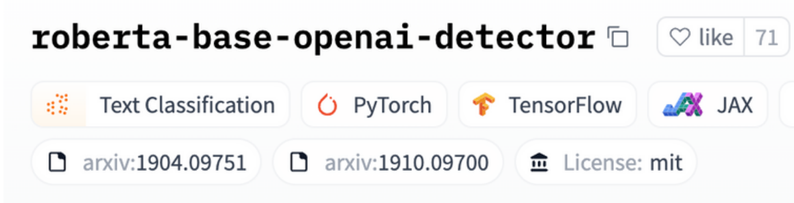
| | id | sum1 | sum2 |
|---|---|---|---|
| 0 | test_0001 | 1 | 0 |
| 1 | test_0002 | 1 | 0 |
| 2 | test_0003 | 0 | 1 |
| 3 | test_0004 | 0 | 1 |
| 4 | test_0005 | 0 | 1 |

# results

| | id | sum1 | sum2 | answer |
|---|---|---|---|---|
| 0 | test_0001 | 1 | 0 | 1 |
| 1 | test_0002 | 1 | 0 | 1 |
| 2 | test_0003 | 0 | 1 | 0 |
| 3 | test_0004 | 0 | 1 | 0 |
| 4 | test_0005 | 0 | 1 | 0 |

1 1

- fine-tuning, basing a sequence classifier on RoBERTaBASE
- RoBERTa is a masked and nongenerative language model that does not share the same architecture or the same tokenizer as GPT-2.
- this model is finetuned for wikipedia and bookcorpus.




- GPTzero is a version of OpenAI's GPT model that has zero pre-existing parameters or weights.
- GPTZero is the most accurate AI detector across use-cases
- GPTZero is finetuned for student writing and academic prose.