SeqXGPT: Sentence-Level AI-Generated Text Detection

Accepted by EMNLP2023

DOI: 10.48550/arXiv.2310.08903

目录

- 一、背景
- 二、相关研究
- 三、本文工作
- 四、实验
- 五、创新思路

一.背景

- LLM的发展和流行可能导致其被滥用

- 文档级检测已不满足检测需求

二.相关研究

- 检测方法

1）仅以文本为数据集的监督式学习

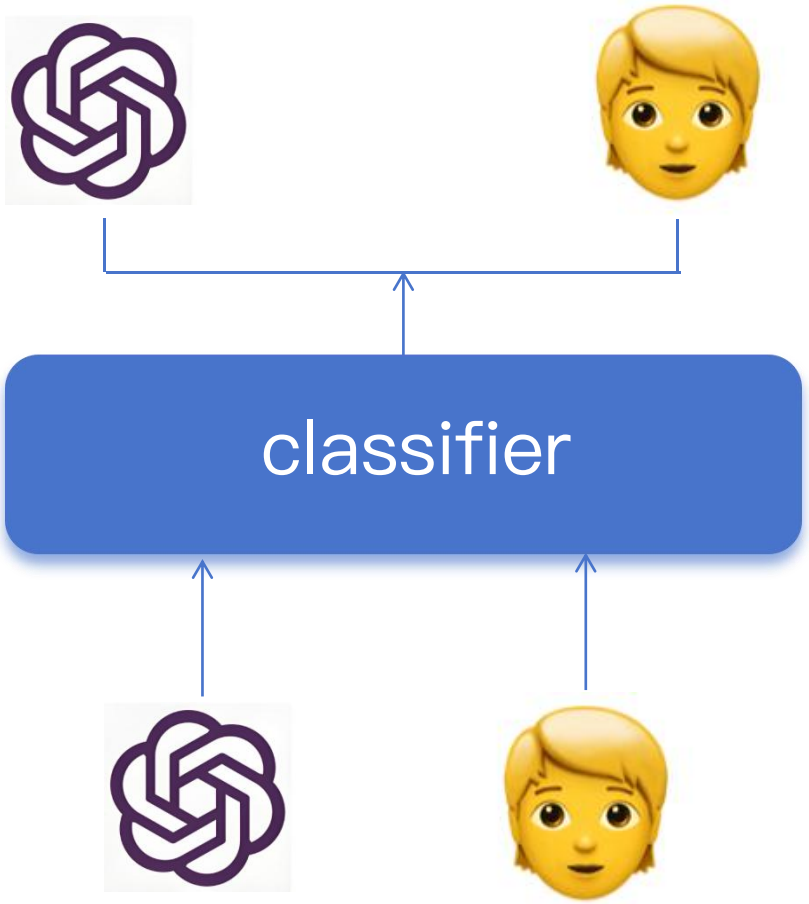
训练数据：[文本,标签]

2）利用模型固有特征

token、token对数概率、token排名.....

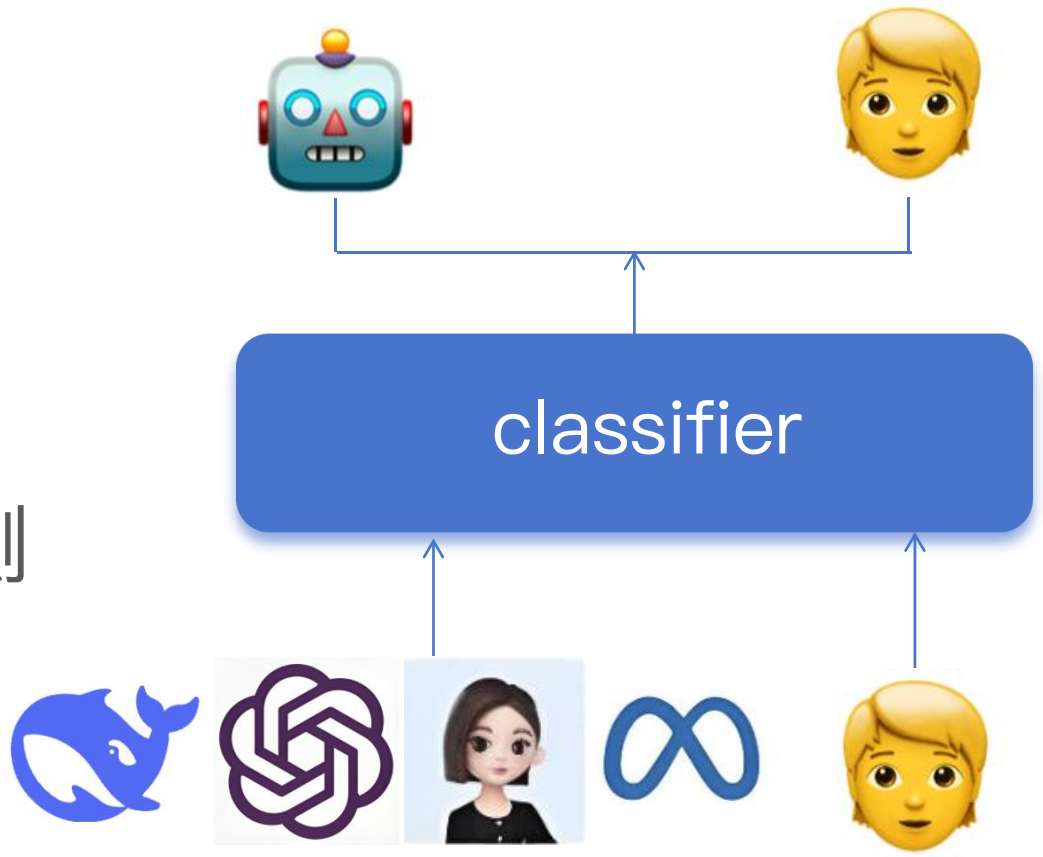
二.相关研究

- 检测任务
  1） 确定模型的二分检测
  2）混合模型的二分检测
  3）混合模型的多类型检测



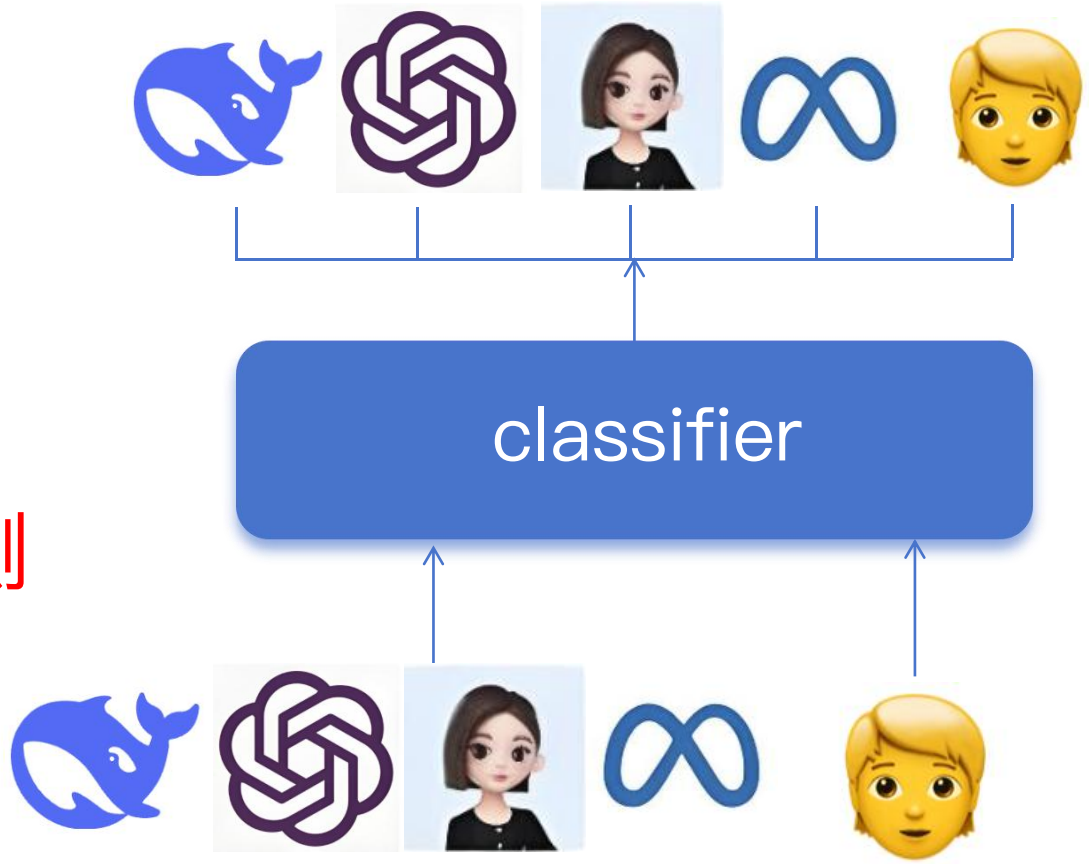
二.相关研究

- 检测任务
  1） 确定模型的二分检测
  2）混合模型的二分检测
  3）混合模型的多类型检测



二.相关研究

- 检测任务
  1） 确定模型的二分检测
  2）混合模型的二分检测
  3）混合模型的多类型检测



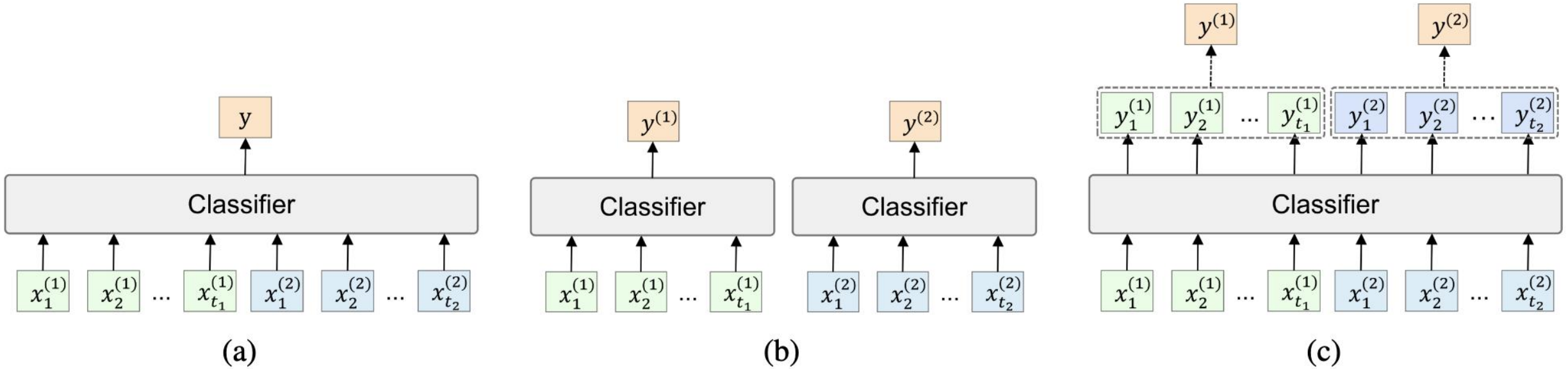
二.相关研究

●检测策略

1） 文档级检测策略

2）句子级检测策略--句子分类

3）句子级检测策略--单词序列标记



三.本文工作

- 1.构建句子级别的检测数据集

- 2.SeqXGPT模型架构

1.构建句子级别的检测数据集

- 文本来源

XSum 的新闻文章

IMDB 的社交媒体帖子、网络文本

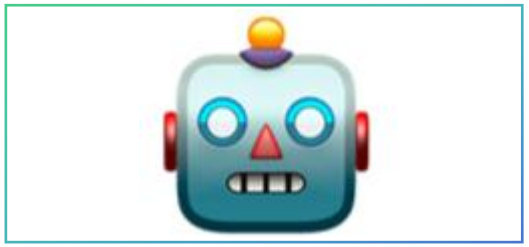
PubMed 和 Arxiv 的科学文章

SQuAD 的技术文档

原始文本数据

每段文字的前1–3句话

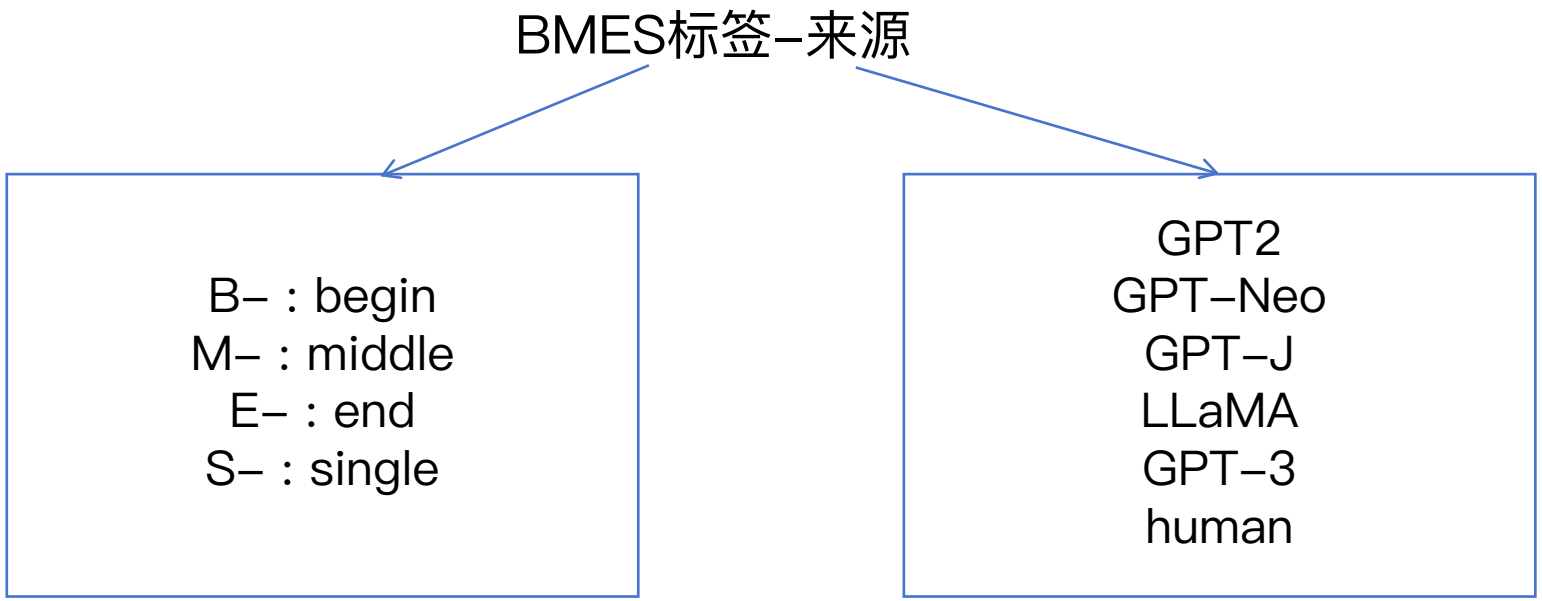
作为提示词



1.构建句子级别的检测数据集

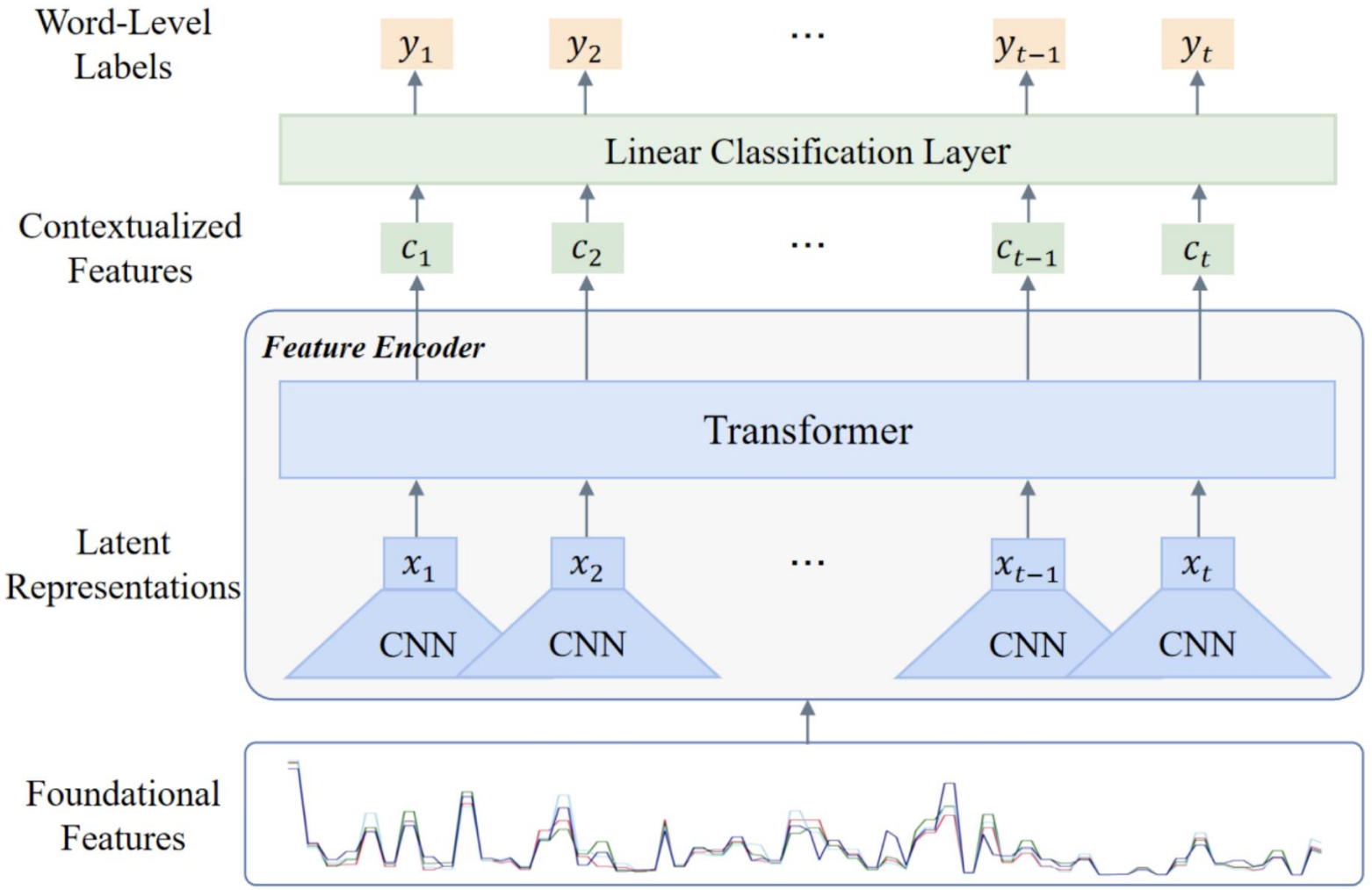
- 标签标注

标签格式：



如：B–human

2.SeqXGPT模型架构



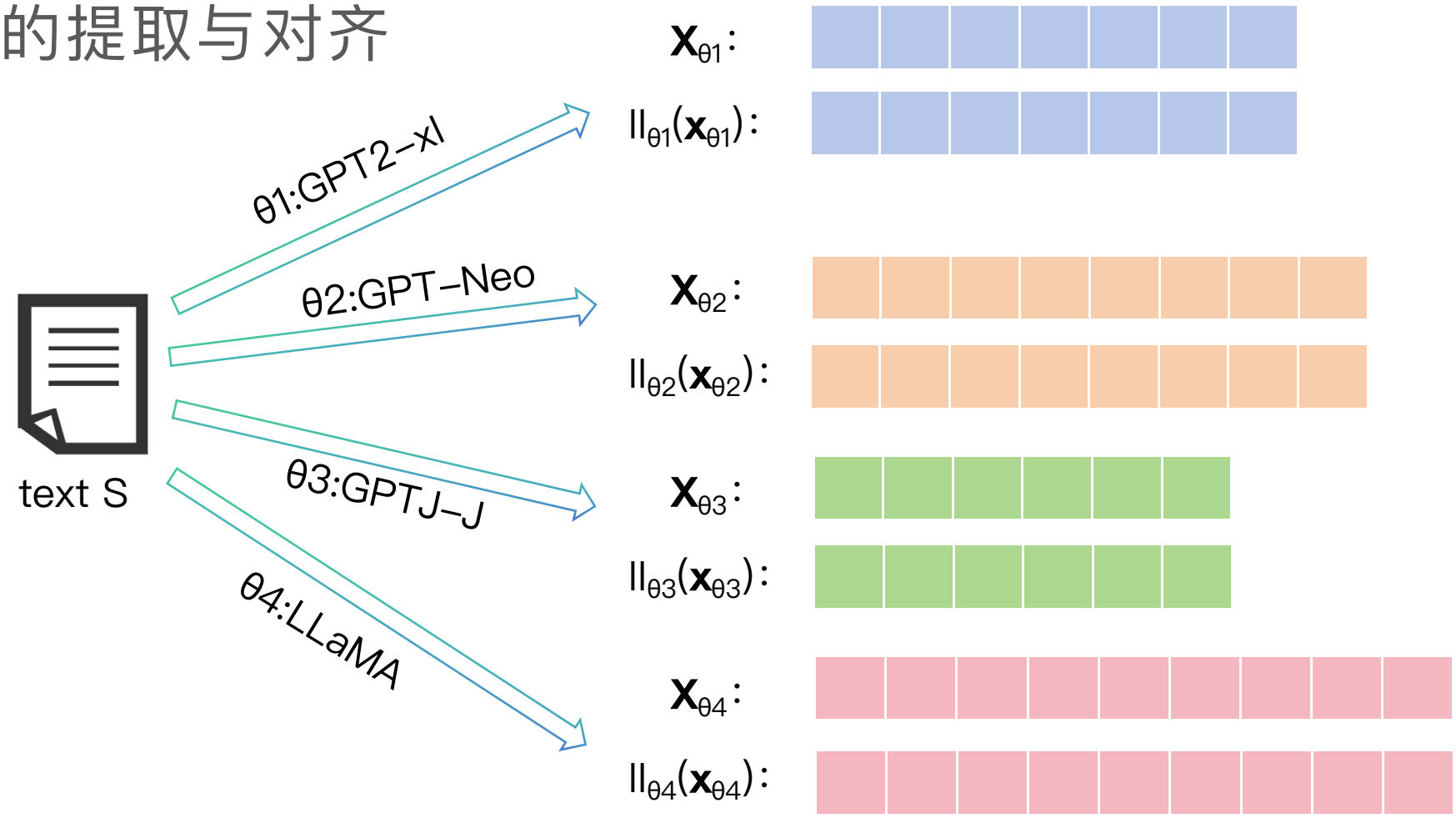
2.SeqXGPT模型架构

- 对数概率的提取与对齐

$$ll_{\theta n}(x_i) = \log p_{\theta n}(x_i|x<i)$$

2.SeqXGPT模型架构

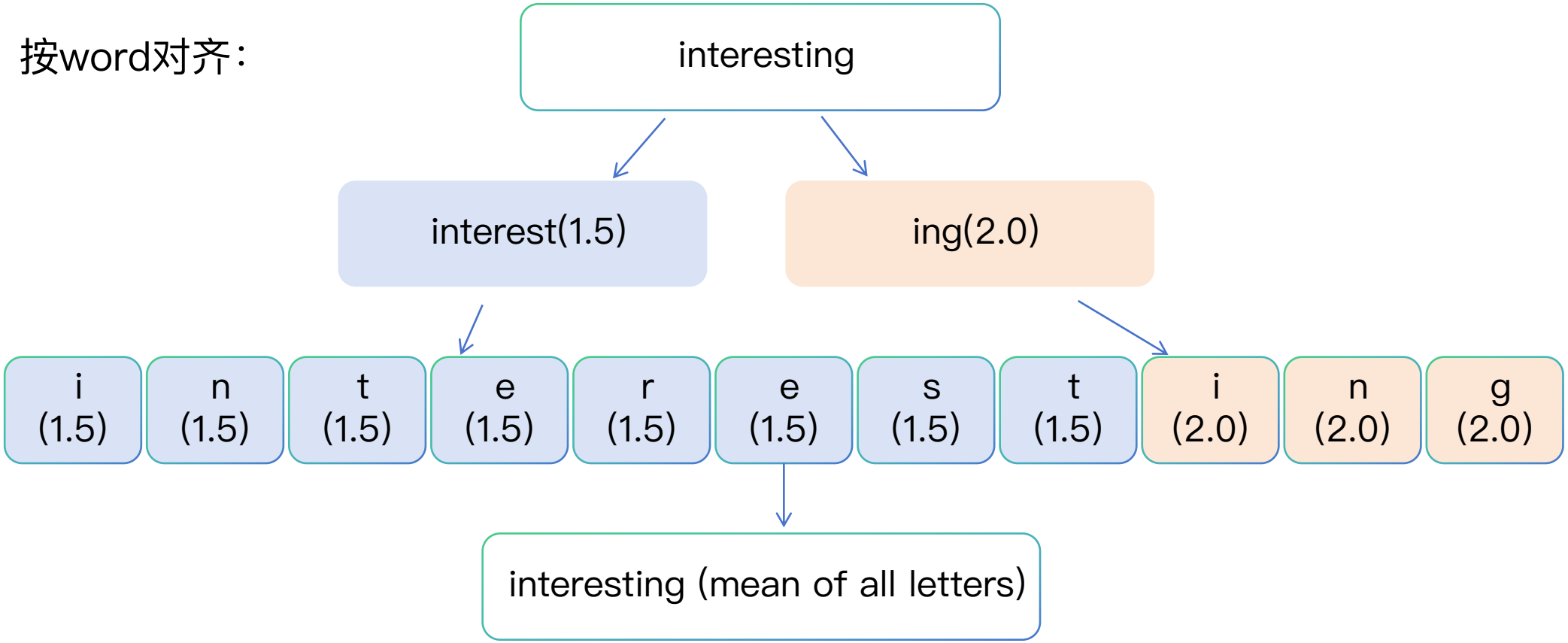
- 对数概率的提取与对齐



2.SeqXGPT模型架构

- 对数概率的提取与对齐

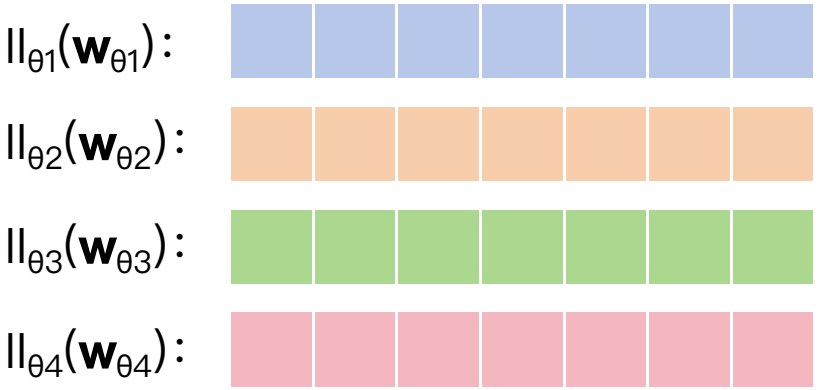
按word对齐：



2.SeqXGPT模型架构

- 对数概率的提取与对齐

结果：



同一段文本在4个白盒模型上获取对数概率
得到4个word级对数概率特征

2.SeqXGPT模型架构

- 特征编码
  - 卷积层

局部的语法或语义结构

不同模型的特征

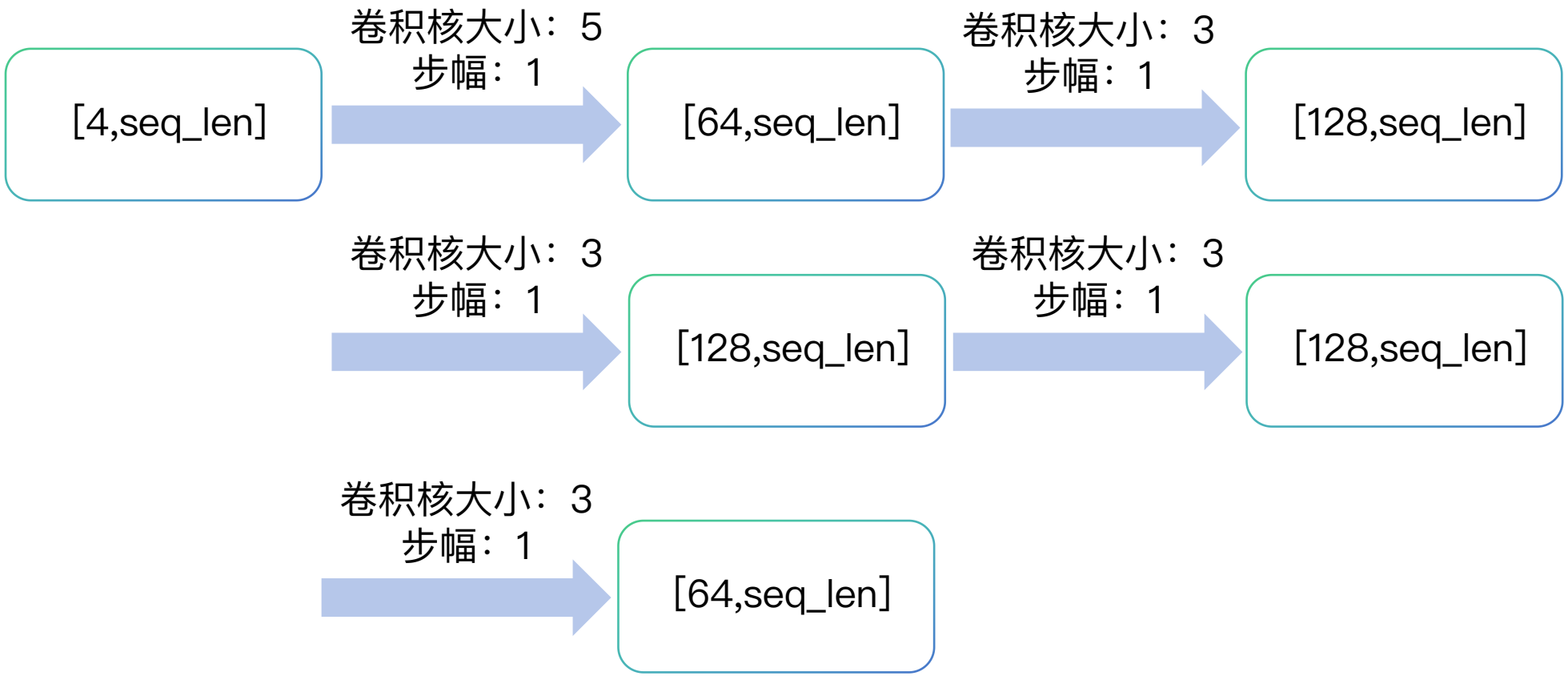
采样随机性而产生的不确定性

声音的结构

不同声源

噪声

●卷积层


[4,seq_len]
卷积核大小：5
步幅：1
[64,seq_len]
卷积核大小：3
步幅：1
[128,seq_len]
卷积核大小：3
步幅：1
[128,seq_len]
卷积核大小：3
步幅：1
[128,seq_len]
卷积核大小：3
步幅：1
[64,seq_len]

2.SeqXGPT模型架构

- 特征编码
  - Transformer层

使用2层Transformer层进行特征编码，提取上下文的特征

输入：CNN层输出的4个[64,seq_len]矩阵
经过转制和拼接后，形成[seq_len,4*64]矩阵
经过两层Transformer encoder
输出：[seq_len,4*64]矩阵

2.SeqXGPT模型架构

- 特征编码
  - 线性分类层

    输入：[seq_len,4*64]的矩阵

    经过全连接神经网络

    输出：24维向量（24个标签）

四.实验

- 白盒模型：

  GPT2-xl (1.5B), GPT-Neo (2.7B), GPT-J (6B) and LLaMA (7B).

- 评估指标：

  Precision (P.)　　:反映准确率

  Recall (R.)　　:反映覆盖率

  Macro-F1 Score　　:结合以上两个衡量指标的综合得分

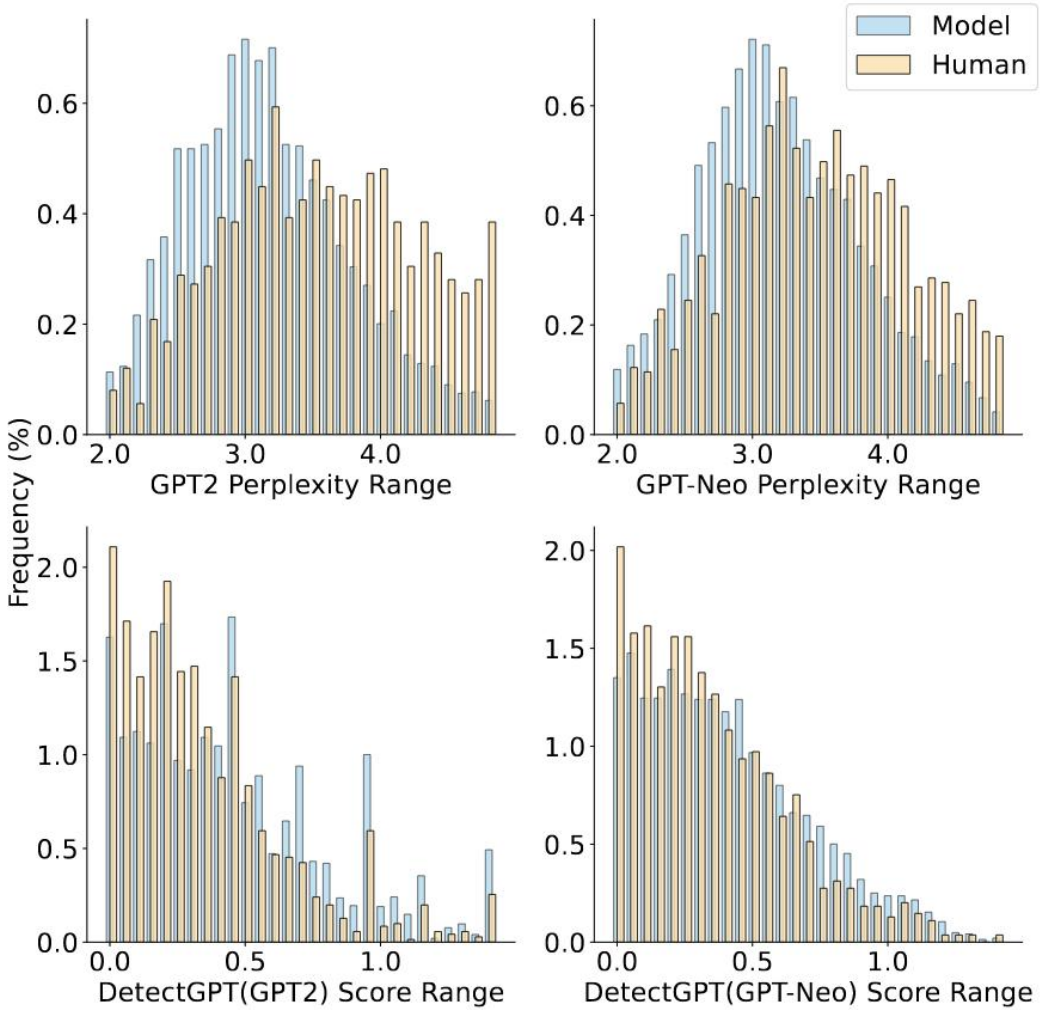
四.实验

- 基线方法
  - 对数概率log p(x)
  - DetectGPT
  - Sniffer
  - RoBERTa
    - -SentRoBERTa
    - -Seq-RoBERTa

[1] MITCHELL E, LEE Y, KHAZATSKY A, 等. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature
[2] LI L, WANG P, REN K, 等. 2023. Origin Tracing and Detecting of LLMs[EB/OL]. arXiv[2025-04-10]. http://arxiv.org/abs/2304.14072.
[3] LIU Y, OTT M, GOYAL N, 等. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach[EB/OL]. arXiv[2025-04-15]. http://arxiv.org/abs/1907.11692.

四.实验

- 句子级检测：


Model
Human
Frequency (%)
GPT2 Perplexity Range
GPT-Neo Perplexity Range
DetectGPT(GPT2) Score Range
DetectGPT(GPT-Neo) Score Range

四.实验

- 句子级检测：

Method	Different AIGT Origins									
	GPT-2					GPT-Neo				
	P.(AI)	R.(AI)	P.(H.)	R.(H.)	**Macro-F1**	P.(AI)	R.(AI)	P.(H.)	R.(H.)	**Macro-F1**
$\log p(x)$	82.2	74.9	43.1	53.9	63.1	81.2	67.8	34.2	51.7	57.5
DetectGPT	80.9	55.4	32.7	62.4	54.3	82.6	44.2	29.1	71.2	49.4
Sent-RoBERTa	89.3	96.9	88.1	66.5	84.4	89.8	95.6	82.7	66.0	83.0
SeqXGPT	**99.3**	**97.9**	**94.5**	**97.1**	**97.2**	**99.5**	**98.2**	**94.8**	**98.1**	**97.6**

基于固定模型的二分检测

四.实验

- 句子级检测：

Method	Different AIGT Origins												Macro-F1
	GPT-2		GPT-2-Neo		GPT-J		LLaMA		GPT-3		Human		
	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	
Sniffer	47.5	56.3	48.4	42.9	39.0	33.5	41.8	16.0	52.8	55.4	51.2	67.2	44.7
Sent-RoBERTa	38.6	48.9	36.9	27.6	34.9	28.7	57.5	33.6	65.5	97.1	89.4	91.6	52.9
Seq-RoBERTa	42.1	81.4	45.3	30.9	61.6	21.6	75.5	82.0	90.3	**98.9**	**94.6**	90.1	64.9
SeqXGPT	**99.2**	**97.9**	**99.3**	**98.2**	**97.6**	**96.8**	**95.8**	**90.8**	**94.1**	93.7	90.7	**95.2**	**95.7**
w/o Transformer	92.4	93.1	92.7	88.9	93.3	62.1	82.1	14.3	22.7	0.2	42.0	95.7	56.9
w/o CNN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.8	100.0	6.6

基于混合模型的多类检测与消融实验

四.实验

- 句子级检测：

Method	Mixed AIGT Origins				
	P.(AI)	R.(AI)	P.(H.)	R.(H.)	Macro-F1
Sniffer	83.2	92.7	67.8	45.3	71.0
Sent-RoBERTa	97.3	97.9	93.5	91.8	95.1
Seq-RoBERTa	96.4	**98.5**	**95.0**	88.9	94.6
SeqXGPT	**98.2**	97.1	91.4	**94.5**	**95.3**

基于混合模型的二分检测

四.实验

- 文档级检测：

Method	Different AIGT Origins												
	GPT-2		GPT-2-Neo		GPT-J		LLaMA		GPT-3		Human		
	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	Macro-F1
Sniffer	76.5	96.6	86.0	83.1	75.0	74.2	92.9	7.0	79.5	83.1	53.4	87.2	67.5
Sent-RoBERTa	45.2	73.0	39.7	46.5	28.3	21.5	72.4	10.5	73.4	**100.0**	**97.4**	92.0	53.4
Seq-RoBERTa	50.4	85.5	42.1	40.0	42.4	26.5	62.6	72.0	85.7	99.0	85.9	36.5	57.9
SeqXGPT	**100.0**	**99.0**	**100.0**	**99.0**	**99.5**	**96.5**	**96.8**	**90.0**	**94.5**	86.5	77.6	**93.5**	**94.2**

基于混合模型的多类检测

四.实验

- 文档级检测：

Method	Different AIGT Origins												Macro-F1
	GPT-2		GPT-2-Neo		GPT-J		LLaMA		GPT-3		Human		
	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	
Sniffer	4.1	80.0	59.9	44.1	21.5	41.2	54.9	14.5	73.0	53.7	35.8	60.0	36.1
Sent-RoBERTa	30.3	35.0	13.6	27.4	24.3	25.3	35.1	27.5	61.7	94.4	75.1	19.1	35.2
Seq-RoBERTa	46.2	64.2	22.7	40.1	60.7	19.8	74.5	75.9	86.2	**99.3**	**89.5**	78.4	60.6
SeqXGPT	**99.5**	**98.4**	**99.1**	**83.6**	**95.5**	**95.0**	**91.6**	**89.1**	**96.5**	91.0	83.1	**94.0**	**92.8**

基于混合模型的多类检测，使用分布外数据集测试

五.创新思路

- 1.情感特征
- 2.信息熵
- 3.使用双向模型
- 4.文本分类&SeqXGPT