基于卷积神经网络的垃圾短信识别app的设计与实现

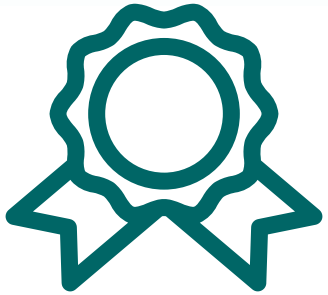
1023040911 蔡启航

目 录

Content

01

研　究　背　景



PART 1 研究背景

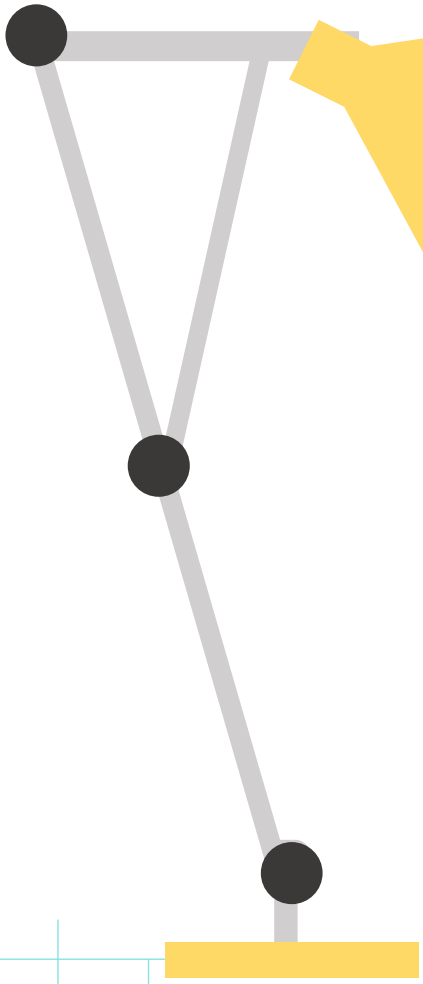
传统解决方案

1、黑白名单过滤
2、基于特征的过滤技术
3、关键字过滤
4、基于内容的识别过滤

缺点与不足

1、不发分子可以通过简单改变短信文本的特征实现短信拦截逃避
2、此类识别过滤技术无法主动适应短信的变化

CNN的优势



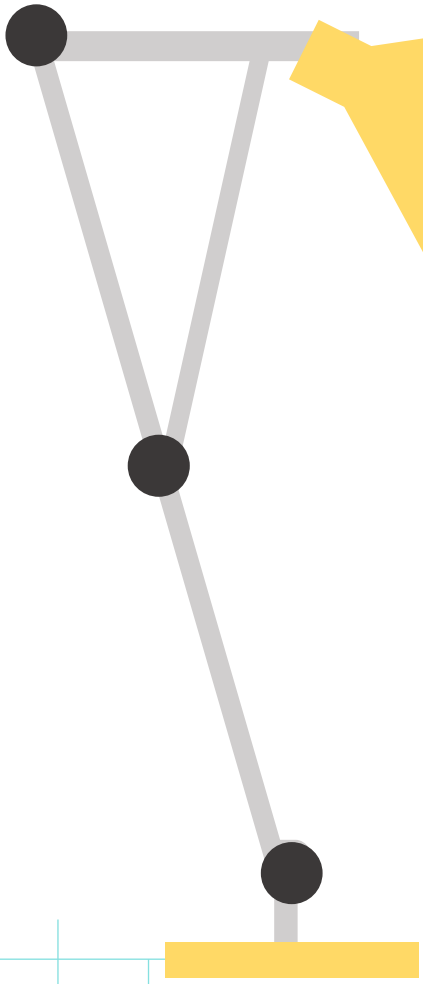
01 局部特征提取：CNN可以通过卷积核在文本中提取局部特征，这些特征可以表示单词、短语或句子的一些基础属性，例如大小写、词性和句法结构等。

02 参数共享：与传统的全连接神经网络不同，CNN的卷积层参数是共享的，这意味着不同位置上的相同特征都将使用相同的参数，减少了需要学习的参数数量，降低了过拟合的风险。

03 并行计算：CNN的卷积操作可以进行并行计算，这样可以加快模型训练的速度，并且在GPU上运行时效率更高。

04 适用于多种文本任务：CNN在文本分类、情感分析、问答系统等多个文本任务中均取得了很好的效果，尤其是在较短的文本序列上表现更为出色。

挑战



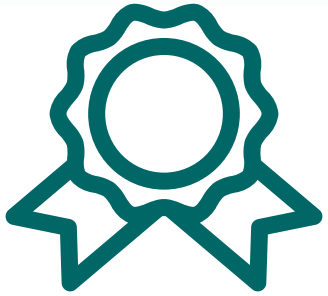
01 短文文本分类算法

02 中文分词方法，文本的传统特征提取方法，分布式特征提取方法

03 如何利用卷积神经网络实现短信文本的预处理

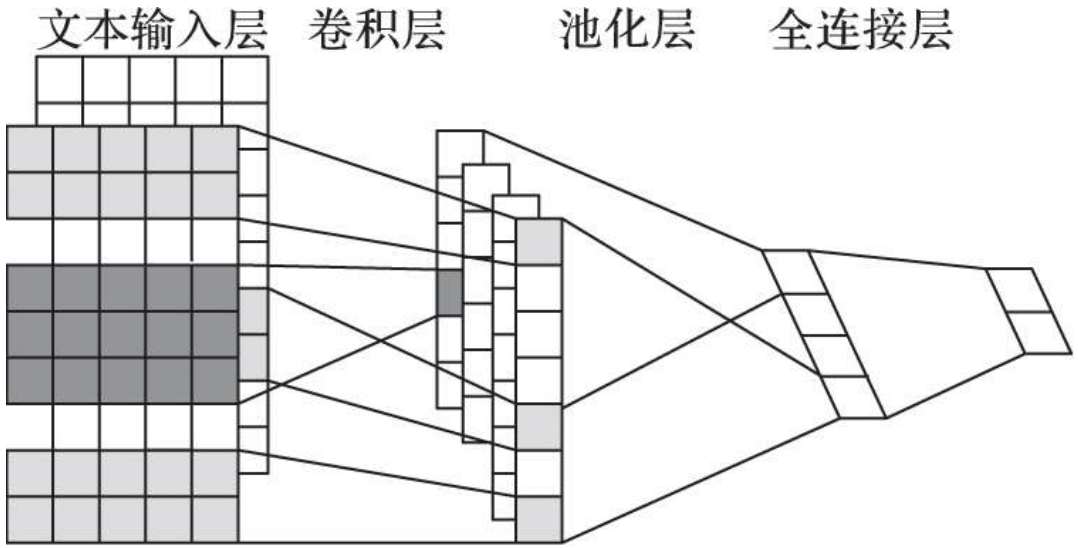
02

主　要　工　作



PART 2 主要工作

1、设计了CNN模型，主要包括：
输入层
卷积层
池化层
全连接层



2、设计并实现了垃圾短信识别算法，并进行了实验测试

PART 2 主要工作

输入层设计：

在CNN中，输入层主要是指用来接收原始短信文本数据的部分。它一般位于神经网络的起始端，作为整个网络的输入。

与传统的全连接神经网络不同，CNN网络通常采用多层卷积层、池化层等结构进行特征提取和降维操作，因此在设计CNN网络时需要对输入层进行一些特殊的处理。

首先，输入层的大小要与输入短信文本的尺寸相对应。

其次，在进行卷积操作前，需要对输入短信文本进行归一化处理。

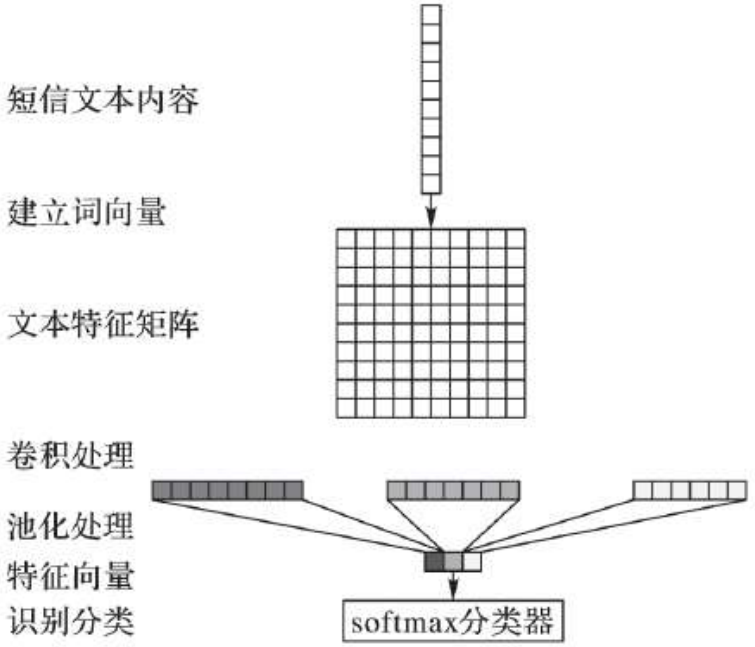
最后，为了防止过拟合，还可以在输入层上添加一些随机噪声或数据增强操作，如旋转、平移、缩放等，以扩充训练数据集，增加模型的泛化能力。

输入层的主要作用是连接CNN模型的下一层，即卷积层，将提取出的表示短信的特征矩阵传递下去。就垃圾短信识别过滤而言，采用CNN模型处理时，模型的输入应该和图像一样为特征矩阵。

PART 2 主要工作

卷积层设计：

CNN中的卷积层是一种用于提取输入短信文本特征的操作。卷积操作的本质是一种特殊的加权求和，其中每个加权项对应着输入短信文本中的一个局部区域的特征值。卷积层包括了一组可学习的卷积核或滤波器，这些卷积核在输入短信文本上进行滑动，每次计算出一个局部区域与该卷积核的卷积结果。通过不断滑动并计算，整张输入短信文本都被处理过，并生成了一个新的特征图。卷积操作可以提取输入短信文本的局部特征，而且同一组卷积核在不同位置能够提取到相似的特征，从而实现了参数共享，减少了模型的参数数量，同时也增强了模型的泛化能力。



PART 2 主要工作

池化层设计：

CNN中的池化层是一种用于减小特征图大小和参数数量的操作。在卷积神经网络中，通过卷积操作可以提取出输入短信文本的特征，这些特征通常是高维的，对应着输入短信文本的不同局部。但是高维的特征会导致模型的参数数量过多，同时也会增加计算量。因此，在卷积神经网络中，需要通过池化层来减小特征图的大小，从而降低参数数量和计算量。

在自然语言处理过程中，需要通过池化层对卷积运算的结果进行局部汇总，以减少从卷积层提取的短信文本向量维数，避免过拟合的发生。常见的池化操作包括Average-Pooling和Max-Pooling。Average-Pooling则是输出特征图V中所有值的平均值。Max-Pooling将特征图V中的最大值输出，这个最大值可以被视为短信的最显著特征。而在大多数自然语言处理任务中，我们常使用Max-Pooling方法。此外，文献[17]也表明Max-Pooling更适用于文本分类，因此本文的研究基于Max-Pooling方法展开。

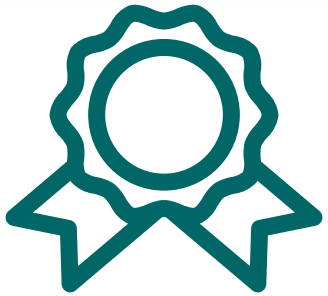
PART 2 主要工作

全连接层设计：

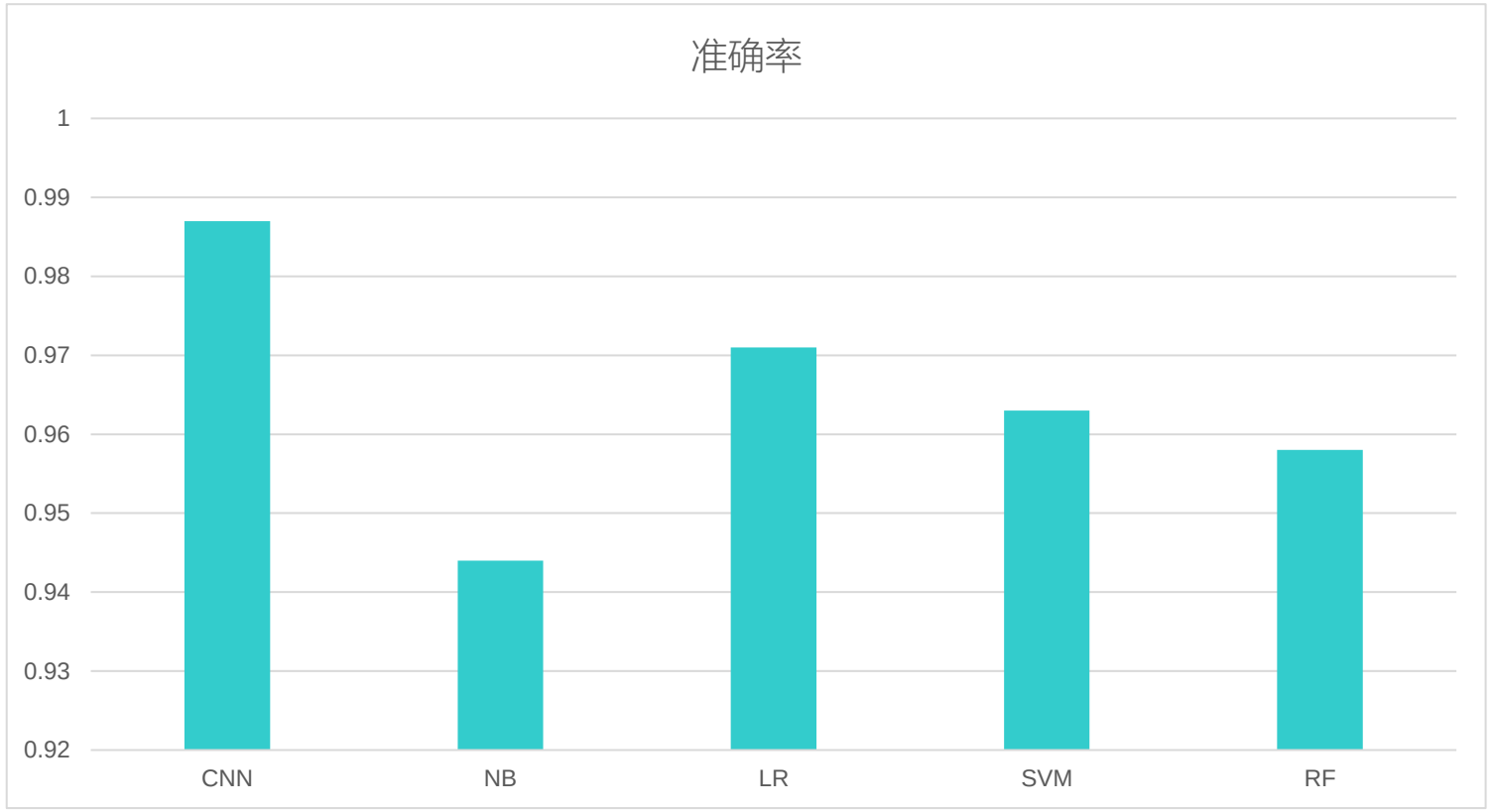
全连接层则是CNN中的一种基础神经网络层。在CNN模型中，全连接层通常被用来将卷积层和池化层提取出来的特征进行分类或回归任务。全连接层的作用是将前面的所有层输出的特征展开成一个向量，然后通过矩阵乘法与权重矩阵相乘，并加上偏置项，得到最终的分类结果。

03

实 验 结 果



PART 3 实验结果



PART 3 实验结果

```
******************** data preprocessing **********************
Building prefix dict from the default dictionary ...
Loading model from cache c:\users\001\appdata\local\temp\jieba.cache
save y successfully!
Loading model cost 0.224 seconds.
Prefix dict has been built succesfully.
******* data preprocessing done in 82.848s *******
```

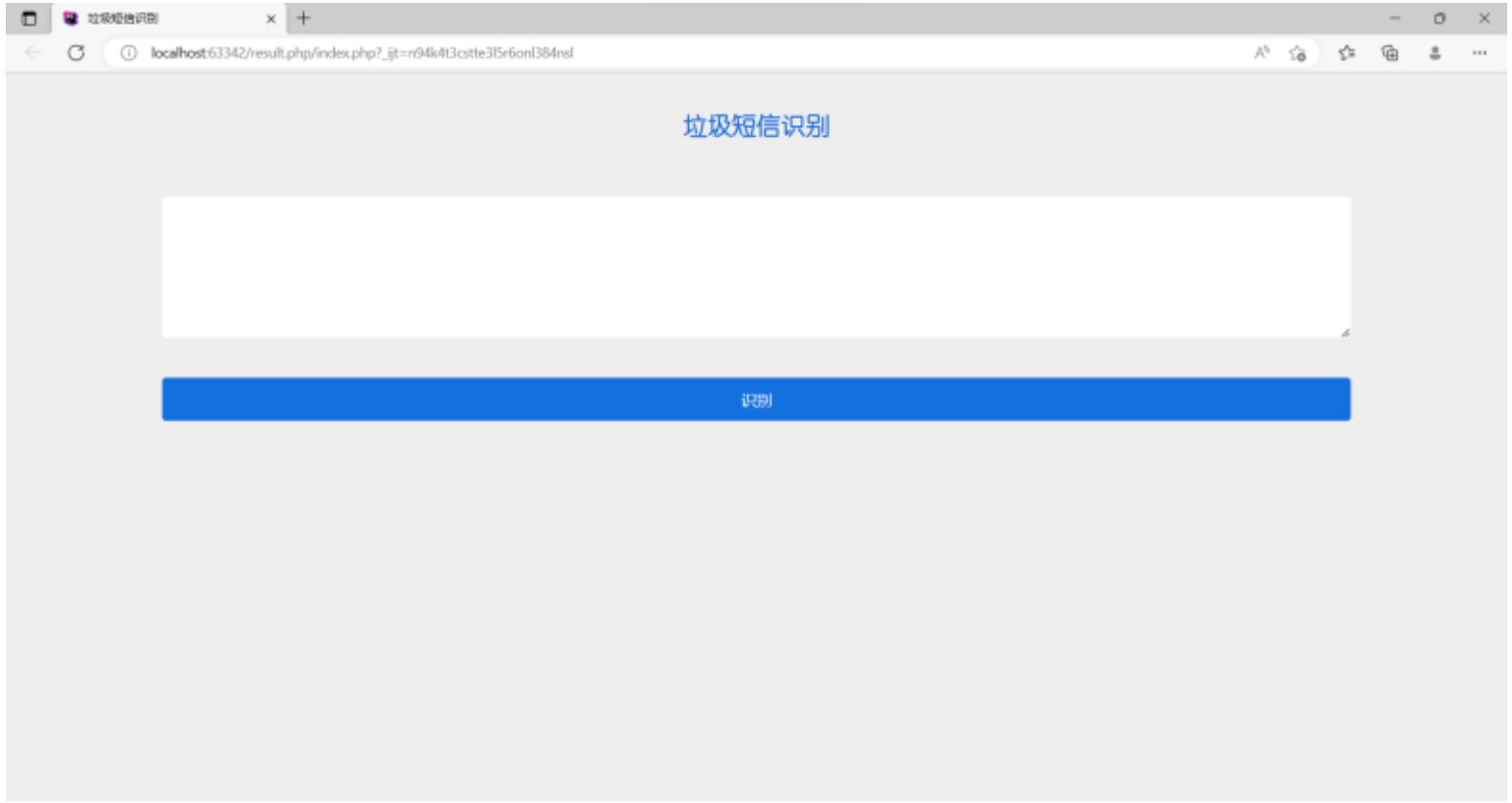
文本预处理阶段控制台信息

PART 3 实验结果

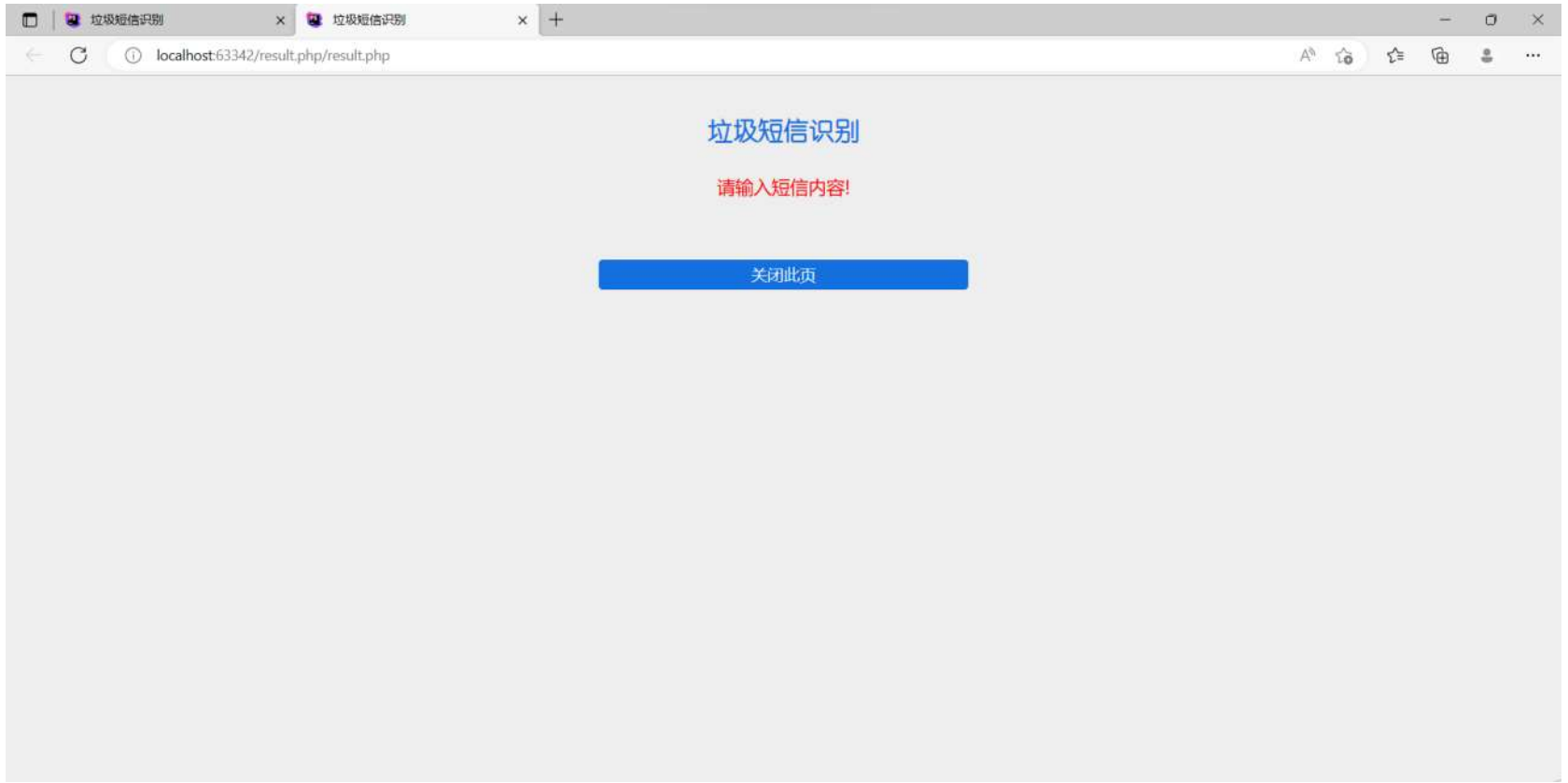
```
******************** CNN ********************
training took 0.023000s!
precision: 93.59%, recall: 93.21%
accuracy: 98.68%
RESULT
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4499
           1       0.94      0.93      0.93       501

   micro avg       0.99      0.99      0.99      5000
   macro avg       0.96      0.96      0.96      5000
weighted avg       0.99      0.99      0.99      5000
```

训练分类器阶段控制台信息

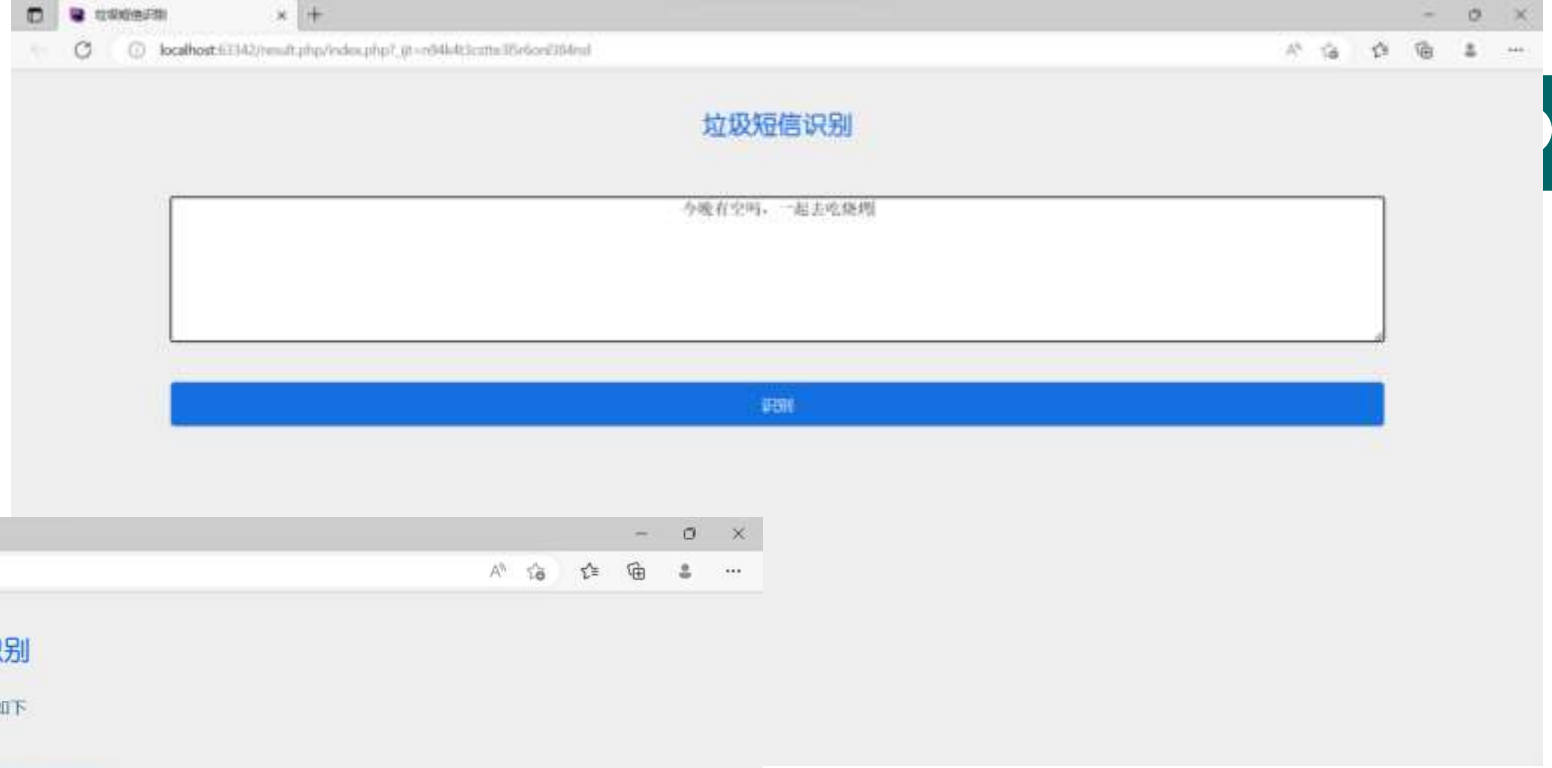
PART 3 实验结果

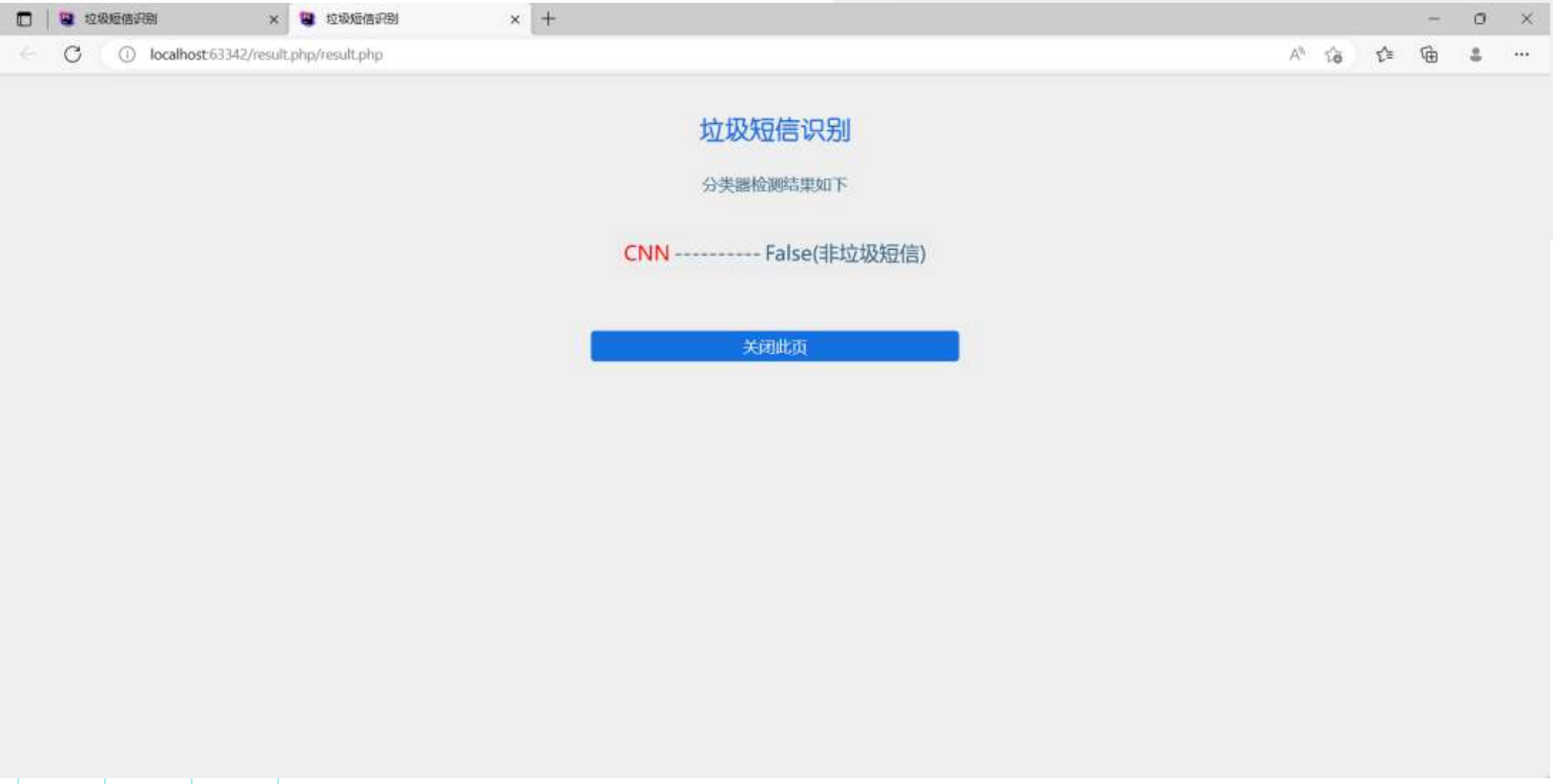

垃圾短信识别
localhost:63342/result.php/index.php?_ijt=n94k4t3cstte3l5r6onl384nsl
垃圾短信识别
识别

PART 3 实验结果

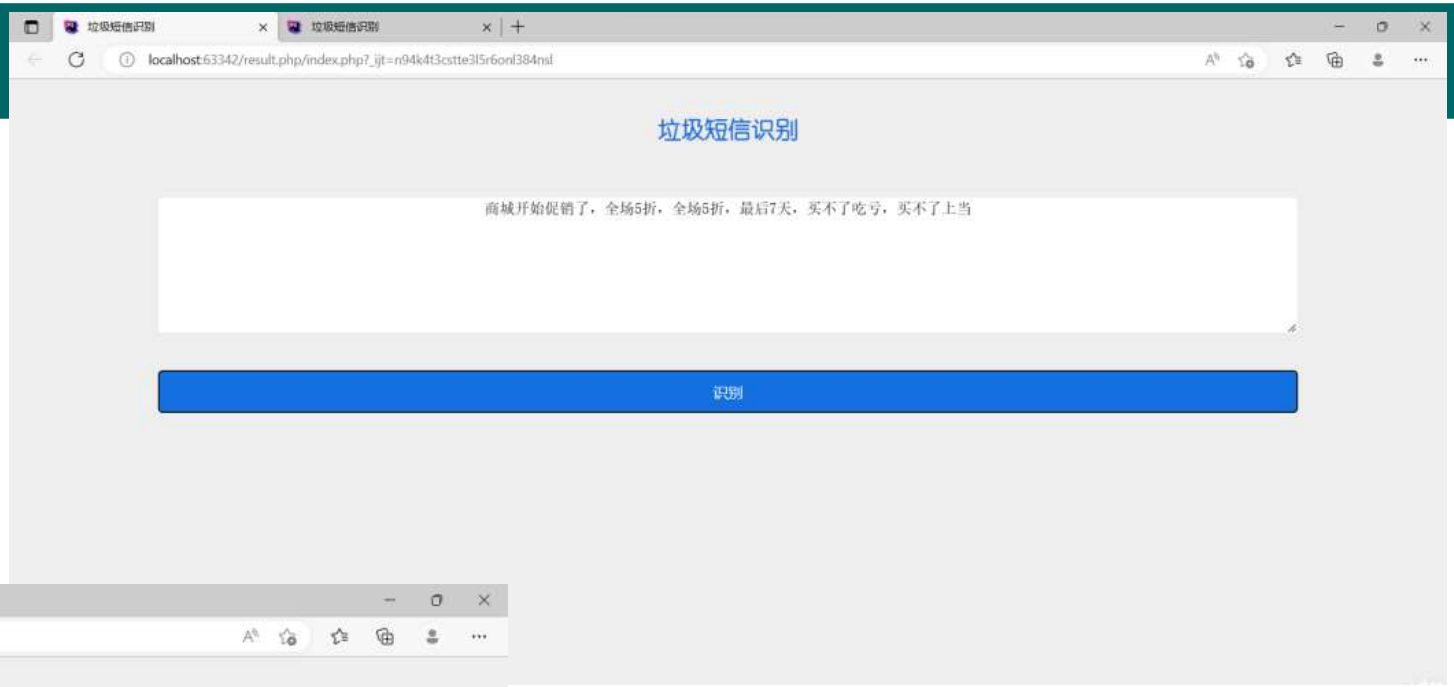

垃圾短信识别
垃圾短信识别
localhost:63342/result.php/result.php
垃圾短信识别
请输入短信内容!
关闭此页

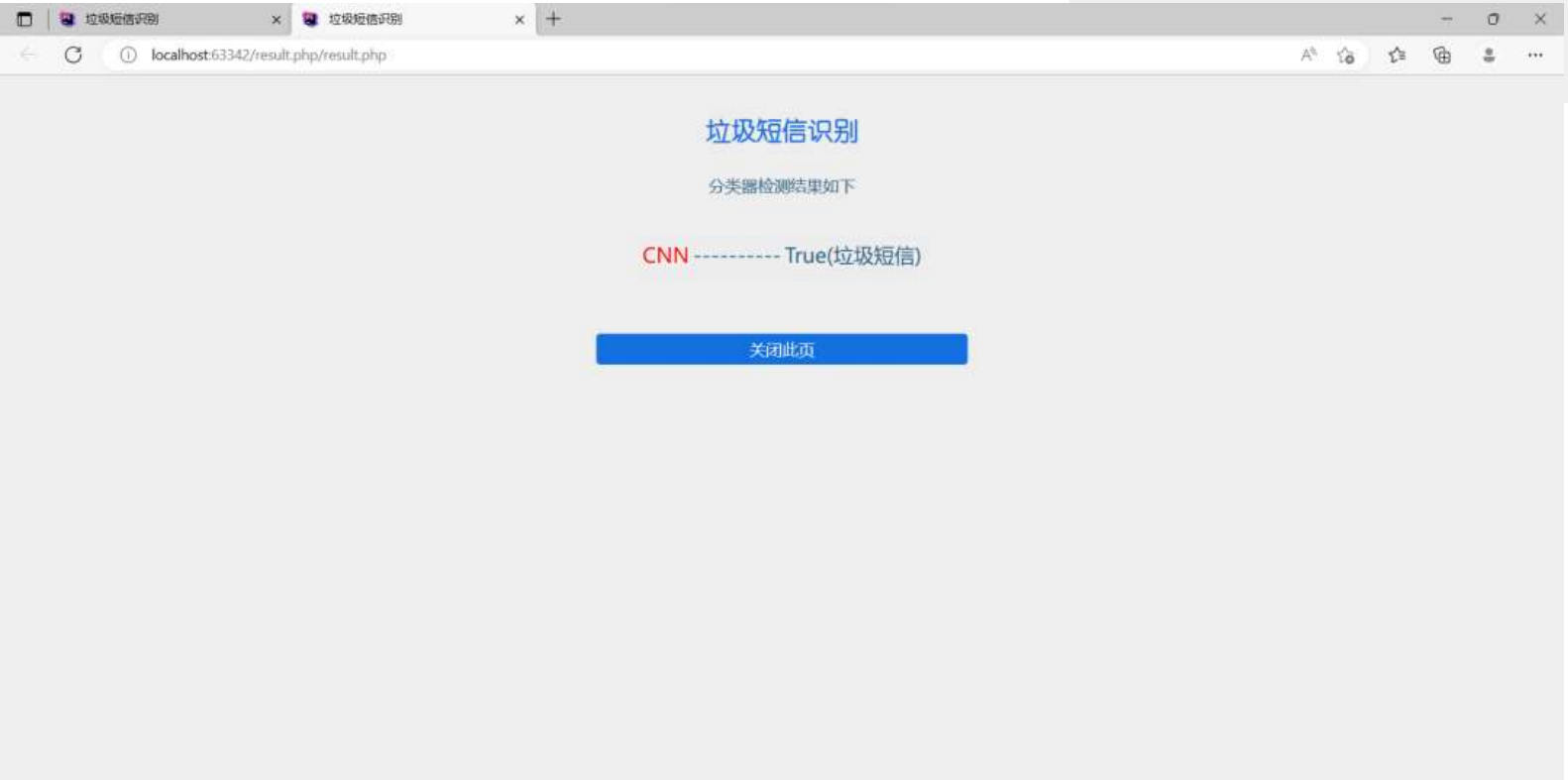
PART 3 实验结果


垃圾短信识别
今晚有空吗，一起去吃烧烤
识别


localhost:63342/result.php/result.php
垃圾短信识别
分类器检测结果如下
CNN ---------- False(非垃圾短信)
关闭此页

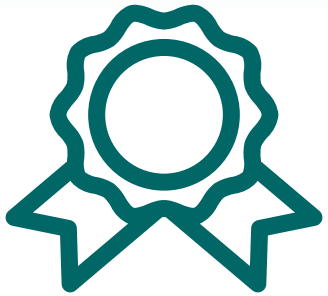
PART 3 实验结果


垃圾短信识别
垃圾短信识别
localhost:63342/result.php/index.php?_ijt=n94k4t3cstte3l5r6onl384nsl
垃圾短信识别
商城开始促销了，全场5折，全场5折，最后7天，买不了吃亏，买不了上当
识别


垃圾短信识别
垃圾短信识别
localhost:63342/result.php/result.php
垃圾短信识别
分类器检测结果如下
CNN ---------- True(垃圾短信)
关闭此页

04

工　作　展　望



PART 4 工作展望

基于卷积神经网络的垃圾短信识别和过滤算法也存在一些需要改进的缺点，例如训练时间长，网络结构中的参数太多。因此，为了提高短信文本识别和分类的效率，缩短训练时间，未来将尝试在分布式平台上测试卷积神经网络的训练结果。

在实际应用中，普通短信的数量往往高于垃圾短信。然而，当前的分类器在设计时假设数据集中普通短信和垃圾邮件的样本分布相对平衡。如果使用这些分类器对具有不平衡类分布的数据集进行分类，将导致分类器性能下降，并且不平衡的数据分布还会引入额外的错误，这些错误可能会对最终的分类结果产生负面影响。因此，下一步的研究重点将是如何识别类别分布不平衡的短信数据集中的垃圾短信。