Advances in Applied Mathematics 应用数学进展, 2022, 11(11), 7681-7687
Published Online November 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.1111813

Hans汉斯

基于TextCNN的涉密文本识别

张　珂[1*]，陈虹瑾[2]

[1]成都信息工程大学网络空间安全学院，四川　成都
[2]四川传媒学院有声语言艺术学院，四川　成都

收稿日期：2022年10月7日；录用日期：2022年11月1日；发布日期：2022年11月10日

摘　要

保密工作直接关系到社会稳定、经济增长、国家安全。新时代信息化和网络迅速普及和发展，办公信息化逐渐成为了主流，在带给办公便利的同时也导致了泄密行为的发生。人工筛选涉密文本极为浪费时间，并且可能会出现人为失误。本文利用爬虫技术构建涉密文本数据集，结合word2vec和TextCNN模型在自建数据集上进行训练。实现准确识别出包含涉密信息的文本。经过实验对比测试，相较于传统的卷积神经网络，TextCNN结合word2vec在自建数据集上达成的效果更好。

关键词

涉密文本，word2vec，TextCNN

Confidential Text Recognition Based on TextCNN

Ke Zhang[1*], Hongjin Chen[2]

[1]School of Cybersecurity, Chengdu University of Information Technology, Chengdu Sichuan
[2]Department of Vocal Language Arts, Sichuan University of Media and Communications, Chengdu Sichuan

Received: Oct. 7th, 2022; accepted: Nov. 1st, 2022; published: Nov. 10th, 2022

Abstract

Confidentiality is directly related to social stability, economic growth and national security. With the rapid popularization and development of informatization and network in the new era, office informatization has gradually become the mainstream, which brings convenience to office work and also leads to the occurrence of leakage of secrets. Manual screening of classified texts is a waste

*通讯作者。

文章引用: 张珂, 陈虹瑾. 基于 TextCNN 的涉密文本识别[J]. 应用数学进展, 2022, 11(11): 7681-7687.
DOI: 10.12677/aam.2022.1111813

of time and may result in human error. In this paper, we use crawler technology to build a secret related text dataset, and combine word2vec and TextCNN models to train on the self built dataset to accurately identify the text containing secret related information. Through experimental comparison test, compared with the traditional convolutional neural network, TextCNN combined with word2vec achieves better results on the self built dataset.

Keywords

Classified Texts, word2vec, TextCNN

1. 引言

在全球信息化快速发展的时代，其已经与各行各业予以结合，达到了工作的便捷性与高效性，多数地方政府已经实现了无纸化电子办公。信息化建设大大地提高了政府的效能，同时也增加了执政的透明度。同时，也让人们对于数据的安全性给予高度重视。依照电子信息系统的实际情况来看，其数据的安全性尤其是涉密信息依然存在一定的漏洞，主要体现在涉密信息无法快速识别、安全等级不清晰、涉密信息监管和处理不当，大多数泄密行为的发生都是因为人员操作不当。快速并准确地查找出包含涉密信息的文本是后续工作地基础。现有的涉密信息检测手段多数是关键词匹配和基于传统机器学习的特征识别，这两种方法局限性较强，只关注了词语和特征本身，忽略了语义的存在。本文提出了结合 word2vec 和 TextCNN 的混合模型，实验结果表明该混合模型能够更准确地判别文本文档中是否存在涉密信息。

2. 相关研究

在之前的一些研究中，卷积神经网络(CNN)取得了不错的效果，但是 CNN 没有考虑文本潜在的主题。

2014 年，Kim [1]提出将 CNN 模型应用到文本分类任务中，发现 TextCNN 模型可以提取文本的语义信息并捕获上下文的相关信息。TextCNN 具有结构简单、训练速度快、效果好等特点。广泛应用于文本分类、推荐等 NLP 领域。

文本循环神经网络(TextRNN)由 Liu，Qiu & Huang [2]提出，与 TextCNN 相比，它可以捕捉文本的时间特征，对文本分类任务有很好的效果，但它的训练速度相对较慢。

曹宇[3]等出了一种基于双向门控循环单元(BiGRUs)的中文文本情感分析方法。首先将文本转换成词向量序列，然后将 BiGRU 用于文本的上下文情感特征。F1 值达到 90.61%，在准确率和训练速度上都高于 CNN 方法。

Chen, Y. [4]等提出了一种基于推特和新浪微博数据的新的情感分析方案，该方案考虑了由表情引起的情感因素。通过参与这些模棱两可的表达，训练出一个情绪分类器，并将其嵌入到基于注意力的长时记忆网络中，对情绪分析有很好的指导作用。

Rehman, A.U. [5]等提出了一种使用 LSTM 和深度 CNN 模型的混合模型。首先使用 word2vec 方法训练初始词嵌入，然后将卷积提取的特征集与具有长期依赖关系的全局最大池层结合起来嵌入后记。该模型还采用了 dropout 技术、归一化和校正线性单元来提高精度。

综上，本文目标是对中文长文本进行分类，综合了上述模型的优势，提出了结合 word2vec 和 TextCNN

的混合模型。经过 word2vec 预处理的中文数据集，再交由 TextCNN 去训练，将很大程度提高使用原模型自带嵌入层的准确率。

3. word2vec 原理介绍

word2vec 最主要的目的就是将不可计算、非结构化的词转换为计算机可以识别出的可计算的向量和结构化的数据。它有两种训练模式，分别是 Skip-gram (跳字模型)和 CBOW (连续词袋模型)。Skip-gram 模型是根据已知中心词语来预测上下文词语；而 CBOW 模型与其相反，是根据已知上下文词语来预测中心词语。Mikilov [6]指出 Skip-gram 相比于 CBOW，虽然训练时间更长，但是在遇到生僻字的时候预测准确率将高于 CBOW。本文将对大量中文文本进行训练，因此选用 Skip-gram 模型。如图 1 所示，Skip-gram 的原理就是用当前中心词是 w_t (爱好)来预测周围的词，将图 1 中窗口大小设为 2，则每次预测 w_t 左边两个词和右边两个词。当输入“爱好”时，输出就会呈现“我”、“的”、“是”、“健身”。

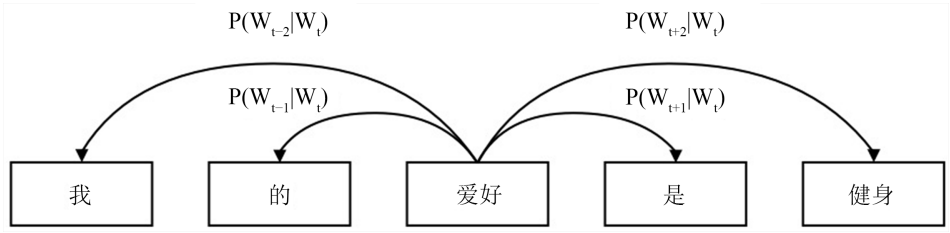

Figure 1. Skip-gram schematic diagram
图 1. Skip-gram 原理示意图

4. 基于 TextCNN 的分类模型设计

TextCNN 是卷积神经网络 CNN 的变型，在网络结构上与传统 CNN 没有变化，包含四个部分：词嵌入、卷积、池化、全连接和 softmax。它通过定义不同大小的过滤核，可以实现提取不同大小的局部特征，从而使得到的特征具有多样性和代表性[7]。TextCNN 目前多用于短文本分类，但它同时也适用于处理长文本[8]。TextCNN 分类模型示意图如图 2。下面具体描述 TextCNN 模型搭建的步骤。

1) 自定义 Embedding 层(嵌入层)

使用 word2vec 模型训练数据集生成相应的词向量，并调整参数，如 embedding_num、batch_size 等，以达到更好的训练结果。将 word2vec 训练好的词向量矩阵作为 TextCNN 的 Embedding 层，这样将会提高模型的准确率。

2) 定义卷积层

本文定义了尺寸分别为 2，3，4 的卷积核，卷积核的数量为 256。使用 ReLU 函数进行激活，使用参数 padding=same 避免卷积时因窗口大小不同输出向量维度不同地现象并确保三个输出。维度与输入维度相同，方便后续操作。使用三种不同尺寸的窗口可以提取到更多的的文本特征信息。也可适当调整超参数，获取更好的训练结果。

3) 定义池化层

TextCNN 输出地向量输入到最大池化层(Max-Pooling)。最大池化层的窗口大小为 2，步长为 2。这样会使最终的词向量维度只减半，上下文关系可以在一定程度上得到保留。

4) 全连接和 softmax

进行最大池化后会得到三个窗口的特征，在全连接层将其连接到一起。Softmax 接收全连接的数据并进行分类。

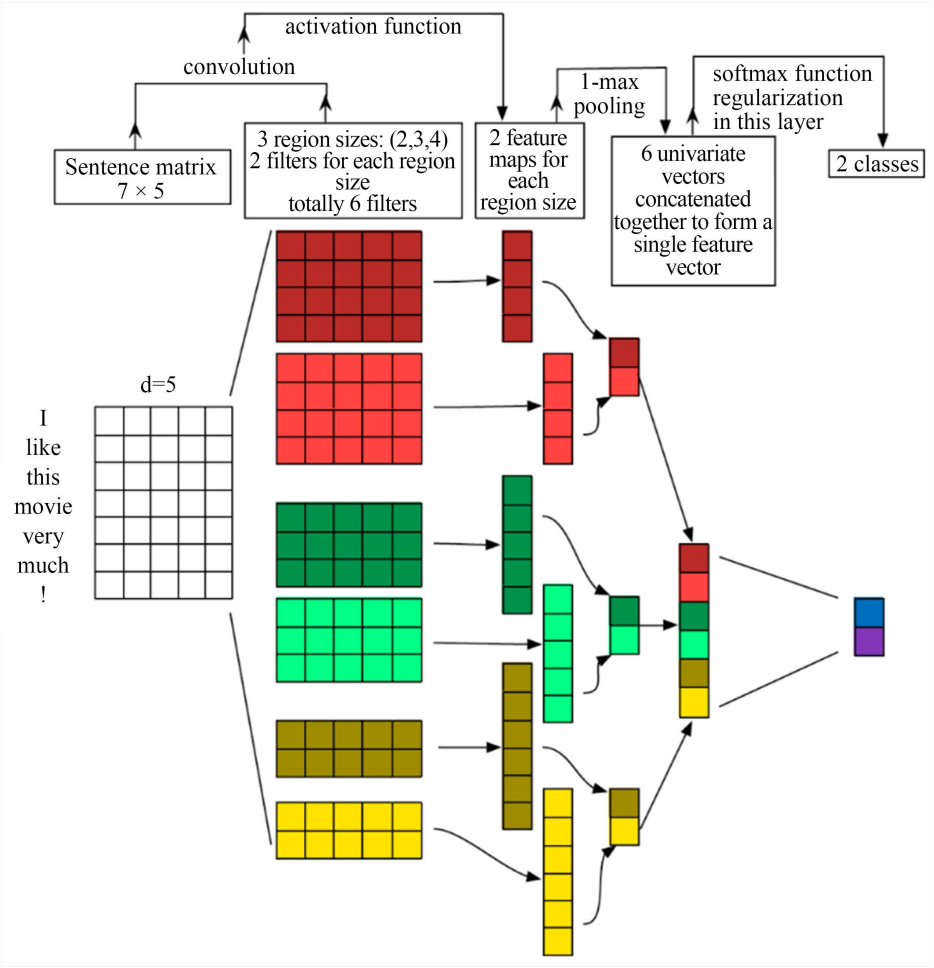

Figure 2. Schematic diagram of TextCNN classification model
图 2. TextCNN 分类模型示意图

在本文的卷积层中，使用不同大小的卷积核。在一个大小为 $n \times h$ 的矩阵中，对于给定的卷积核 $\omega \in R^{hk}$ 和一个大小为 $x_{i:i+h-1}$ 的窗口进行卷积操作生产特征函数 $c_i = f\left(\omega \cdot x_{i:i+h-1} + b\right)$。在这个函数中，$x_{i:i+h-1}$ 表示一个大小为 $h \times k$ 的窗口，由输入矩阵 i 到 $i+h-1$ 行组成，由 $x_i, x_{i+1}, \cdots, x_{i+h-1}$ 拼接而成。h 代表窗口中的词数，ω 是权重矩阵，b 是偏执参数，f 是非线性函数。每个卷积操作意味着一个特征向量的提取。通过定义不同大小的窗口可以提取到不同的特征向量，形成卷积层的输出。本文在池化层中选取最大池化方式，从每个滑动窗口生成的特征向量中选择一个最大特征，然后将其连接起来用向量表示。使用 softmax 函数在最终的全连接层进行分类，其公式为：

$$P_i = \frac{e^i}{\sum^j e^i}$$

其中 P_i 表示第 i 类的概率，e^i 表示第 i 类输出的对应值，j 表示类的总数。

5. 实验结果与分析

5.1. 数据集获取

涉密文件是指以文字、图片、音像及其他记录形式记载商业、国家秘密内容的资料。本文针对以文

字为记录形式的涉密文件即涉密文本进行研究。

由于该研究的特殊性，并无现成数据集用于训练。因此，本数据集通过爬虫爬取维基解密上一些已解密的中文文本，主要针对军事和政治两类具有鲜明特征的涉密文本。数据集中还包含同类型的(政治、军事)非涉密文本和其他类型的非涉密文本。非涉密文本是采用了搜狗实验室提供的搜狐新闻数据集[9]。

5.2. 文本预处理

构建完成的数据集不能直接用于训练模型，需要对数据集进行预处理，主要包含中文文本分词、去停用词[10]。本文选用 jieba 中文文本分词工具进行分词，并导入去停用词表完成去停用词。

文本预处理后还需要对文本标签进行分类，本实验数据集中的文本标签只有两类，即涉密文本和非涉密文本。对于二分类任务可以使用 one-hot (独热编码)处理文本标签数据。处理完成后为涉密文本的文本转化为[0,1]，非涉密文本转化为[1,0] [11]。处理完成后就可用于模型的训练。

5.3. 实验环境及模型参数

1) 实验环境

本文所用的实验环境如表 1 所示。

Table 1. Experimental environment
表 1. 实验环境

CPU	AMD Ryzen 7 5800H
GPU	NVIDIA GeForce RTX 3070
内存	16GB
操作系统	Windows 11
编程语言	Python 3.7

2) 模型参数

本文配置的模型参数如表 2 所示。

Table 2. Model parameter configuration
表 2. 模型参数配置

TextCNN	
Parameters	Values
Convolution kernels number	256
Convolution kernels size	2,3,4
Activation function	ReLU
Pooling strategy	1-max pooling
Dropout	0.5
L2 regularization	3

5.4. 评价标准

为了评价模型的预测性能，本文选取准确率(Accuracy)、损失率(Loss)两个指标作为模型的评测指标，假设 TP 为真阳性，FP 为假阳性，TN 为真阴性，FN 为假阴性。模型的准确率为准确预测样本数/样本总

数[4]。这就引出了定义公式：

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

选择交叉熵损失函数，其公式如下：

$$C=-\frac{1}{n}\sum_{x}\left[y\ln a+(1-y)\ln(1-a)\right]$$

其中 x 表示样本，y 表示标签，a 表示预测的结果，n 表示样本总量。

5.5. 实验及结果

为了体现该模型的准确率，分别采用了 TextRNN、TextRNN + Attention、TextRCNN 三种模型在相同数据集上进行分类。经过对比，TextCNN 在本数据集上的表现更好。不同模型的准确率如表 3 所示。

Table 3. Comparison of model accuracy
表 3. 模型准确率比较

模型名称	准确率/%
TextCNN	93.54%
TextRNN	92.12%
TextRNN + Attention	90.90%
TextRCNN	91.54%

6. 结语

本文先构建了一个包含涉密中文文本的数据集，利用 word2vec 对数据预处理后，基于 TextCNN 模型进行了分类任务。通过改进后的 TextCNN 模型，经过实验对比，在自建数据集上取得了更好的分类效果。

本文也存在一些不足，比如用来训练模型的数据集数据量不够，模型也并未做出太大的调整。下一步将继续研究是否有更优秀的模型，在保证训练速度的同时也能够能达到更好的分类效果。

参考文献

[1] Kim, Y. (2019) Convolutional Neural Networks for Sentence Classification. arXiv preprint arXiv:1408.5882.

[2] Cambria, E., Liu, Q., Decherchi, S., *et al*. (2022) SenticNet 7: A Commonsense-Based Neurosymbolic AI Framework for Explainable Sentiment Analysis. *Proceedings of LREC* 2022.

[3] 曹宇, 李天瑞, 贾真, 殷成凤. BGRU: 中文文本情感分析的新方法[J]. 计算机科学与探索, 2019, 13(6): 973-981.

[4] Chen, Y., Yuan, J., You, Q., *et al*. (2018) Twitter Sentiment Analysis via Bi-Sense Emoji Embedding and Attention-Based LSTM. P*roceedings of the* 26*th ACM international conference on Multimedia*, 117-125.

[5] Rehman, A.U., Malik, A.K., Raza, B., *et al.* (2019) A Hybrid CNN-LSTM Model for Improving Accuracy of Movie Reviews Sentiment Analysis. *Multimedia Tools and Applications*, **78**, 26597-26613.

[6] Mikolov, T., Chen, K., Corrado, G., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv e-prints.

[7] 李志杰, 耿朝阳, 宋鹏. LSTM-TextCNN 联合模型的短文本分类研究[J]. 西安工业大学学报, 2020, 40(3): 6.

[8] 李悦, 汤鲲. 基于 TextCNN 的政策文本分类[J]. 电子设计工程, 2022(12): 43-47.

[9] 于海. 基于卷积神经网络的非结构化文本敏感信息检测系统的设计与实现[D]: [硕士学位论文]. 北京: 北京邮电大学, 2019.

[10] 刘春磊, 武佳琪, 檀亚宁. 基于 TextCNN 的用户评论情感极性判别[J]. 电子世界, 2019(3): 2.

[11] 张浩然, 谢云熙, 张艳荣. 基于 TextCNN 的文本情感分类系统[J]. 哈尔滨商业大学学报(自然科学版), 2022(3): 285-292.