

基于注意力机制的社交垃圾文本检测方法

曲强, 于洪涛, 黄瑞阳

(国家数字交换系统工程技术研究中心, 河南 郑州 450002)

摘要: 在社交网络中, 大量的垃圾文本严重威胁用户的信息安全与社交网站的信用体系。针对噪声性与稀疏性问题, 提出一种基于注意力机制的卷积神经网络检测方法。在经典卷积神经网络的基础上, 该方法增加了过滤层, 并在过滤层设计基于朴素贝叶斯权重技术的注意力机制, 解决了噪声性问题。并且, 它改变了池化层原有的策略, 采用基于注意力机制的池化策略, 缓解了稀疏性问题。结果表明, 相对于其他检测方法, 所提方法的检测准确率在4个数据集上分别提高了1.32%、2.15%、0.07%、1.63%。

关键词: 社交网络; 信息安全; 垃圾文本; 注意力机制

中图分类号: TP309

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020002

Attention-based approach of detecting spam in social networks

QU Qiang, YU Hongtao, HUANG Ruiyang

National Digital Switching System Engineering & Technological R&D Center, Zhengzhou 450002, China

Abstract: In social networks, a large amount of spam has seriously threaten users' information security and the credit system of social websites. Aiming at the noise and sparsity problems, an attention-based CNN method was proposed to detect spam. On the basis of classical CNN, this method added a filter layer in which an attention mechanism based on Naive Bayesian weighting technology was designed to solve the noise issue. What's more, instead of the original pooling strategy, it adapted an attention-based pooling policy to alleviate the sparsity problem. Compared with other methods, the results show that the accuracy has increased by 1.32%, 2.15%, 0.07%, 1.63% on four different data sets.

Key words: social networks, information security, spam, attention system

收稿日期: 2019-03-25; 修回日期: 2019-07-11

通信作者: 于洪涛, yht_ndsc@139.com

基金项目: 国家自然科学基金创新群体基金资助项目 (No.61521003)

Foundation Item: The National Natural Science Foundation Innovation Group Project (No.61521003)

论文引用格式: 曲强, 于洪涛, 黄瑞阳. 基于注意力机制的社交垃圾文本检测方法[J]. 网络与信息安全学报, 2020, 6(1): 54-61.

QU Q, YU H T, HUANG R Y. Attention-based approach of detecting spam in social networks[J]. Chinese Journal of Network and Information Security, 2020, 6(1): 54-61.

1 引言

随着移动互联网技术的广泛应用, 在线社交网络由于具有便捷、灵活、内涵丰富的特性而快速成为人们生活重要的组成部分, 如 Facebook、Twitter、Google、新浪微博、微信等流行社交网络。目前, 在线社交网络的用户数量呈指数级别增长, 据统计, 2018 年春节期间微信和 WeChat 的合并月活跃账户数超过 10 亿。社交网络由于蕴含的巨大用户隐私信息及其广阔的商业价值, 成为不法分子图谋不轨的目标。大量发送垃圾文本是不法分子攻击社交网络的重要手段之一。其中, 本文中的垃圾文本指: 以为商品营造虚假事实、威胁网络安全等为目的, 大量发送的商品推销文本、虚假评论文本、热点事件的谣言等文本内容。根据 2013 年的社交网络垃圾文本统计报告, 2013 年 1-6 月, 社交垃圾文本数量增长 355%, 每 200 条社交文本中有 1 条是垃圾文本, 它们对 5% 的社交应用 App 造成一定程度的威胁。社交网络的垃圾文本不仅恶化社交网络环境, 影响用户体验, 危害用户信息安全, 而且对社交网络的可用性以及安全性造成一定程度的影响^[1]。

针对社交网络垃圾文本的危害, 国内外学者对社交网络垃圾文本检测方法进行了大量的理论研究。Ratkiewicz 等^[2-3]提出基于关键词的检测方法, 该方法利用统计的方法提取垃圾文本中频繁出现的关键词, 并根据这些关键词检测垃圾文本。该方法计算复杂度较低, 并且取得了一定的检测效果, 但存在以下两个问题: ① 该方法需要人工设计提取的关键词, 不具有普适性; ② 该方法提取关键词特征表示, 在计算机中表示为 0/1, 并且其维度大小为关键词数目, 因此关键词特征表示存在稀疏性问题^[4]。

为解决基于关键词方法存在的非普适性问题与稀疏性问题, 研究者提出基于神经网络的检测方法。例如, CNN^[5]、RNN^[6]、LSTM^[7]、GRU^[8]等神经网络方法。其中, CNN 模型利用卷积核与池化核可以有效提取文本的深层认知特征; RNN 等模型利用时序神经单元提取文本的时序特征。但它们都存在以下 2 个问题: ① 该方法对原始文本进行检测, 缺乏合理的处理文本噪声单词的机

制; ② 该方法池化层策略虽然选取了具有显著区分能力的特征, 如最大池化策略选取最突出的特征, 平均池化策略选取平均的特征等, 但这些池化策略目的单一, 并且不能动态地进行优化, 即不能自动选取贡献最大的文本特征。

为实现自动选取贡献最大的文本特征, 研究人员提出注意力机制模型, 即根据句子中每个单词对于分类结果的贡献程度, 赋予每个单词不同的权重, 以此表示模型对于每个单词的注意力。目前, 基于注意力机制的深度神经网络方法根据利用信息类型的不同可以分为两类: 利用外部信息的方法和不利用外部信息的方法。① 在利用外部信息的方法中, AHNN^[9]、AP-CNN 方法^[10]利用 CNN 模型提取文本整体含义, 辅助 LSTM 模型进行注意力表示, 但模型中的 CNN 模型仅起到辅助生成作用, 并且使用文本信息的外部信息帮助, 计算复杂度较高。② 在不利用外部信息的方法中, Semantic-CNN 方法^[11]在过滤层加入注意力机制, 提出利用语义特征进行初始化过滤, 来提升文本分类的实验精度。AP-BiRNN^[12]、AP-BiLSTM 方法^[13]在特征表示层与分类层间加入注意力机制, 进一步识别有效特征单词, 提升了文本分类精度。虽然基于注意力机制的神经网络方法利用注意力机制可以提取足够的特征表示, 缓解稀疏性, 但是依旧存在噪声性^[4]问题。

针对基于关键词方法面临的稀疏性问题, 基于神经网络与注意力机制方法面临的噪声性问题, 在不借助外部信息的情况下, 本文提出基于注意力机制的卷积神经网络 (MA-CNN) 检测方法, 主要内容如下。

1) MA-CNN 在原始 CNN 模型基础上增加过滤层, 并且在过滤层融合关键词提取的思想, 设计基于朴素贝叶斯权重技术^[14]的注意力机制。通过提取具有检测效果的关键词, 该方法有效降低了文本中噪声含量, 解决了噪声性问题。

2) MA-CNN 在池化层融合注意力机制的思想, 采用基于注意力机制的池化策略。通过注意力机制给具有检测效果的单词赋予较高的权重, 有效提取文本的特征表示, 缓解稀疏性问题。需要说明的是, 本文中的注意力机制池化策略与目前流行的 RNN 注意力机制不同, 不同之处在于

RNN 注意力机制是基于每个时刻单元重新赋予权重,而本文中的注意力机制池化策略是基于每个核内部单元重新赋予权重。

3) 最后,本文在 Youtube、Opinion、Telephone 与 E-mail 这 4 个数据集上进行实验,结果表明,相比于传统深度神经网络方法与基于注意力机制的深度神经网络方法,本文提出方法的检测准确率在 4 个数据集上分别提高了 1.32%、2.15%、0.07%、1.63%。

2 MA-CNN 模型

针对社交网络文本的噪声性与稀疏性问题,本文提出基于注意力机制的卷积神经网络模型,该模型分为 5 层,分别是过滤层、嵌入层、卷积层、池化层以及分类层。

2.1 过滤层

针对文本内容包含大量噪声的问题,在过滤层,MA-CNN 模型基于关键词的检测方法的思想,对某条社交网络信息包含的单词序列 $[x_1, x_2, \dots, x_n]$,设计基于朴素贝叶斯权重技术的注意力机制。

基于朴素贝叶斯权重技术的注意力机制:根据式(1)计算每个单词的朴素贝叶斯权重,然后按照条件选取一定数量的关键词,从而过滤噪声。

$$r_i = \frac{(p_s^{x_i} + \alpha) / \|p_s\|_1}{(p_{\tilde{s}}^{x_i} + \alpha) / \|p_{\tilde{s}}\|_1} \quad (1)$$

其中, $p_s^{x_i}$ 表示在垃圾信息类 s 的文本中包含词 x_i 的文本数目, $p_{\tilde{s}}^{x_i}$ 表示在非垃圾信息类 $\tilde{s}$ 文本中包含词 x_i 的文本数目, $\|p_s\|_1$ 表示垃圾信息类 s 的文本数目, $\|p_{\tilde{s}}\|_1$ 表示非垃圾信息类 $\tilde{s}$ 的文本数目, α 是平滑参数。文中选取 r_i 大于 1 的单词进入嵌入层。

2.2 嵌入层

经过过滤层的注意力机制提取关键词之后,假设某条社交网络信息由 n 个单词过滤到 k 个单词,输入嵌入层,即 $[x_1, x_2, \dots, x_k]$ 。在嵌入层,根据不同的表示方法构建嵌入矩阵,每个单词 x_i 被映射成一个实数域的特征向量 $e_i \in R^{d \times 1}$ 。具体过程包含两个步骤。

(1) 构建嵌入矩阵 $W^{wrd} \in R^{d \times |V|}$, 其中, V 是一个固定大小的词典集合, d 是嵌入单词的维度,

嵌入矩阵 W^{wrd} 需要利用不同的表示方法进行学习,文中分别采用了随机嵌入、SkipGram^[15]、CBOW^[15]与 Glove^[16]等 4 种特征表示方法。其中, SkipGram 方法与 CBOW 方法都属于 Word2vec 模型,类别上属于基于预测的词向量模型,但有一定的区别: SkipGram 方法通过目标单词来预测上下文语境单词的方式获取每个单词的特征向量表示,而 CBOW 方法通过上下文语境单词预测目标单词的方式获取每个单词的特征向量表示,这类基于预测的词向量模型缺点在于其对统计信息利用得不充分,训练时间与语料大小息息相关,优点在于可以提取词与词之间的线性关系。而 Glove 方法通过在共现矩阵上构造类似于线性关系的条件,不仅提高了训练速度,而且利用了语料库的全局统计信息。

(2) 对于每个单词 x_i , 通过矩阵乘法将单词 x_i 转换成嵌入表示 e_i , 表达式如式(2)所示。

$$e_i = W^{wrd} v^i \quad (2)$$

其中, $v^i \in R^{|V| \times 1}$, 在 e_i 的指标位置为 1, 其余位置为 0。

最终,嵌入层形成某条社交网络信息的特征向量表示为 $[e_1, e_2, \dots, e_k]$ 。

2.3 卷积层

2014 年, YoonKim 提出新的 CNN 模型^[5], 解决了文本分类问题并且取得了良好的效果。本文提出的 MA-CNN 模型在卷积层沿用了 YoonKim 设计模型的卷积核结构,但在池化层放弃了原来的最大池化策略,采用新的注意力机制池化策略,图 1 展示了仅带有大小为 2 卷积核的 MA-CNN 模型的卷积层、池化层以及分类层设计。

如果某条社交网络信息的特征向量表示为 $[e_1, e_2, \dots, e_k]$, 那么利用 $e_{1:k} = e_1 \oplus e_2 \oplus \dots \oplus e_k$ 表示某条社交网络信息的特征表示,同理, $e_{i:i+j}$ 表示包含单词的这段文本的特征表示, $e_{i:i+j} \in R^{jd \times 1}$ 。假设卷积层包含 m 种卷积核结构,大小为 $h_i, i \in \{1, \dots, m\}$, 每种卷积核数目为 M 个,那么第 i 种第 j 个卷积核关于第 1 个单词的输出如式(3)所示。

$$c_{i,j,l} = \sigma(w_i e_{l:l+h_i-1} + b_i) \quad (3)$$

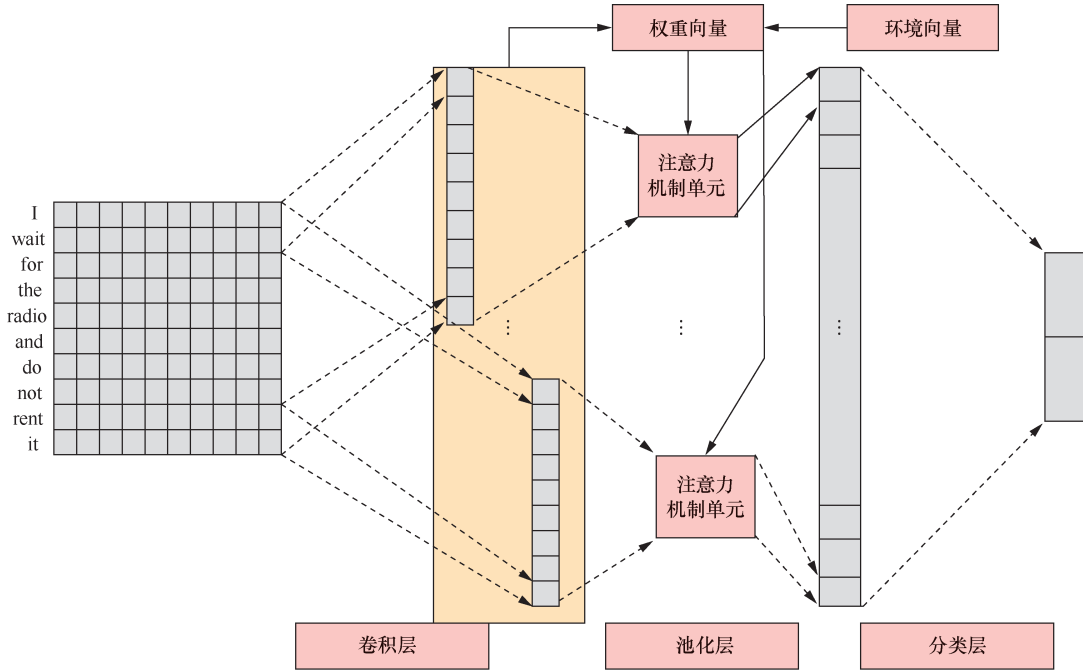


图 1 卷积核为 2 的 MA-CNN 模型

Figure 1 The MA-CNN model with convolution kernel size of two

其中, $w_i \in R^{1 \times h_d}$, $b_i \in R$ 分别表示第 i 种卷积核的权重矩阵与偏置, $\sigma(\bullet)$ 表示激活函数。如果某条社交文本信息含有 k 个单词并且采用第 i 种卷积核结构, 那么该条社交文本信息可以被分割为 $\{e_{1:h_i}, e_{2:h_i+1}, \dots, e_{k-h_i+1:k}\}$, 第 i 种第 j 个卷积核的输出为 $\vec{c}_{i,j} = \{c_{i,j,1}, c_{i,j,2}, \dots, c_{i,j,k-h_i+1}\}$, $\vec{c}_{i,j} \in R^{(k-h_i+1) \times 1}$ 。

2.4 池化层

针对 CNN 池化策略单一并且不能动态更新的问题, MA-CNN 模型在池化层放弃了原来的最大池化策略, 根据注意力机制的思想, 提出基于注意力机制的池化策略。

基于注意力机制的池化策略: 根据第 i 种卷积核结构的所有输出, 可以得到第 i 种卷积核结构的输出矩阵 $C_i = \{\vec{c}_{i,1}, \vec{c}_{i,2}, \dots, \vec{c}_{i,M}\}$, $C_i \in R^{(k-h_i+1) \times M}$ 。根据注意力机制的思想, MA-CNN 模型动态重新更新注意力机制池化策略的权重, 处理过程如式(4)所示。

$$\begin{cases} V_i = \tanh(W^{(v)} C_i^T) \\ a_i = \text{soft max}(V_i u_e) \\ s_i = C_i^T a_i \end{cases} \quad (4)$$

其中, $W^{(v)}$ 表示注意力机制池化策略矩阵, V_i 表

示第 i 种卷积核结构更新后的输出。 u_e 表示环境向量, 这里处理方式随机化选取, 并且在学习过程中动态更新。 a_i 表示第 i 种卷积核结构的注意力机制池化策略的权重, s_i 表示池化层第 i 种卷积核结构输出。 $W^{(v)} \in R^{(k-h_i+1) \times M}$, $V_i \in R^{(k-h_i+1) \times (k-h_i+1)}$, $u_e \in R^{(k-h_i+1) \times 1}$, $a_i \in R^{(k-h_i+1) \times 1}$, $s_i \in R^{M \times 1}$ 。由于 MA-CNN 采用 m 种卷积核结构, 那么某条社交网络信息的特征表示为 $S = s_1 \oplus s_2 \oplus \dots \oplus s_m$, $S \in R^{mM \times 1}$ 。

2.5 分类层

根据句子的特征表示 S , 文中使用广泛的 soft max 函数进行完成分类任务, 即检测社交网络垃圾信息任务, 处理过程如式(5)所示, 其中, $y^s \in R^{2 \times 1}$ 表示样本 S 的预测概率分布, $\hat{y}^s \in R^{2 \times 1}$ 表示样本 S 的真实概率分布。 $W^{(p)}, b^{(p)}$ 表示预测概率的矩阵与偏置, $W^{(p)} \in R^{2 \times mM}$, $b^{(p)} \in R^{2 \times 1}$ 。

$$y^s = \text{soft max}(W^{(p)} S + b^{(p)}) \quad (5)$$

基于交叉熵的定义, 本文中 Loss 函数定义如式(6)所示, 根据损失函数的梯度求导, 可以进行反向传播学习。

$$\text{Loss}(\theta) = - \sum_{i=1}^2 \hat{y}_i^s \log(y_i^s) \quad (6)$$

3 实验

3.1 实验设置

(1) 数据集

为验证 MA-CNN 方法的检测效果,本文通过在 4 个数据集上进行实验,进一步评价 MA-CNN 的检测效果。本文使用的 4 个数据集情况如表 1 所示。

表 1 实验数据集
Table 1 Experimental data sets

数据集	垃圾文本的检测准确率	正常文本的检测准确率	文本总数
Youtube	1 005(51.25%)	956(48.75%)	1 961
Opinion	800(50%)	800(50%)	1 600
SMS	3 198(13.41%)	4 827(86.59%)	5 574
E-mail	1 813(39.40%)	2 788(60.60%)	4 601

1) Youtube 数据集

该数据集是利用 Youtube 的 API 爬取形成的,包含 2015 年 1—3 月的 5 个访问次数最多的视频的评论数据,共 1 005 条垃圾评论信息和 956 条正常评论信息。

2) Opinion 数据集

该数据集由几个网站的数据集组成,包含 400 条来自 TripAdvisor 网站的可信积极评论,400 条来自 Expedia、Hotels.com、Orbitz、Priceline、TripAdvisor、Yelp 等网站的可信消极评论,400 条来自 Amazon Mechanical Turk 的虚假积极评论,400 条来自 Amazon Mechanical Turk 的虚假消极评论。

3) SMS 数据集

该数据集由 4 个 SMS 数据集组成,包含来自 Grumbletext Web Site 的 425 条垃圾信息,来自 NUS SMS 中心的 3 375 条正常信息,来自 Caroline Tag's PhD 论文的 450 条正常信息,来自 SMS Corpus 0.1 的 1 002 条正常信息和 433 条垃圾信息。

4) E-mail 数据集

该数据集通过邮件系统采集形成,包含 1 813 条垃圾邮件信息与 2 788 条正常邮件信息。

(2) 实验环境

本文使用的实验环境为 Ubuntu16.04 系统,

8 个 CPU 处理器以及 23.5 GB 内存,编程语言为 Python3.6,编程框架为 Tensorflow。

(3) 实验设置

在过滤层,MA-CNN 使用基于朴素贝叶斯权重技术的注意力机制。在嵌入层,MA-CNN 使用随机化的表示方法,单词表示向量维度为 128 维。在卷积层,MA-CNN 使用核大小为 3、4、5 的 3 种卷积核,每种卷积核数目为 100 个。在池化层,MA-CNN 采用基于注意力机制的池化策略。在分类层,MA-CNN 采用 softmax 函数。在实验中,训练集与测试集的比例为 9:1,实验过程进行 10 次,取 10 次检测准确率的均值作为最终的结果。

(4) 评价指标

本文采用检测准确率评价实验各个方法的实验效果,这里的检测准确率是所有社交文本的检测准确率,不仅包含垃圾文本的检测准确率,而且包含正常文本的检测准确率。因此,检测准确率的计算方法如式(7)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

其中,TP 表示真实类别为正常文本,模型预测类别仍为正常文本的文本数目;TN 表示真实类别为垃圾文本,模型预测类别仍为垃圾文本的文本数目;FP 表示真实类别为垃圾文本,模型预测类别却为正常文本的文本数目;FN 表示真实类别为正常文本,模型预测类别仍为垃圾文本的文本数目。

3.2 实验结果

在实验部分中,本文首先将提出的 MA-CNN 方法与 6 种方法进行对比,包含 4 种传统的深度神经网络方法与 2 种目前流行的基于注意力机制的深度神经网络方法 (semantic-CNN、AP-BiLSTM),验证 MA-CNN 方法的检测性能效果。其次,在 CNN 模型的基础上,分别对嵌入层表示方法、卷积层激活函数以及池化层注意力机制对检测结果的影响进行了探讨。

实验 1 不同检测方法的检测性能对比

根据表 2 的实验结果,在 4 个经典的深度神经网络算法中,CNN 模型在 Youtube、Opinion 与 SMS 这 3 个数据集上的检测准确率比较高,而 LSTM 模型在 E-mail 数据集上的检测准确率比较

高,可以看出 CNN 模型在文本分类与检测的效果一般好于循环神经网络模型。因为卷积神经网络更加注重文本内容的深度语义特征,而循环神经网络更加注重文本字符的时间序列特征,所以对于文本分类与检测这种偏向于区分语义特征的任务而言,卷积神经网络,即 CNN 模型的检测效果表现得更好。由于 CNN 模型良好的检测效果,因此在后续对于各个层次的机制与方法的实验中,实验选取的基础模型是 CNN 模型,不是循环神经网络模型。

表 2 不同检测方法的检测性能对比
Table 2 The detection performance comparison of different detection methods

检测方法	Youtube	Opinion	SMS	E-mail
CNN ^[5]	92.86	86.26	99.64	87.17
RNN ^[6]	87.24	70.62	98.74	85.65
LSTM ^[7]	89.80	78.75	98.20	89.78
GRU ^[8]	87.75	81.25	98.92	89.56
Semantic-CNN ^[11]	95.11	90.30	99.82	87.14
AP-BiLSTM ^[13]	91.35	91.16	99.75	88.61
本文方法(MA-CNN)	96.43	93.31	99.89	91.41

另外,将 6 种方法在各个数据集上检测效果比较好的方法与本文提出的 MA-CNN 模型进行对比,相对于每个数据集上的最佳算法,MA-CNN 模型在 4 个数据集上分别有 1.32%、2.15%、0.07%、1.63%的提升,验证了 MA-CNN 模型可以有效检测社交网络垃圾文本。

实验 2 嵌入层表示方法的影响

在 CNN 模型的基础上,本部分实验由嵌入层表示方法对检测准确率的影响进行探究,分别采用 4 种表示方法:随机化方法、SkipGram 方法、CBOW 方法与 Glove 方法。

根据表 3 的实验结果,使用不同的表示方法的 CNN 模型在不同数据集上表现性能不一般。例如,对于 SMS 数据集,使用 Random+CNN 模型的检测准确率最高;对于 Youtube 与 E-mail 数据集,使用 CBOW+CNN 模型的检测准确率最高;对于 Opinion 数据集,使用 Glove+CNN 模型的检测准确率最高。尽管对于这 4 个数据集,

使用 SkipGram+CNN 模型的检测准确率都不是最高,但检测准确率稳定在比较高的水平。

表 3 嵌入层表示方法的影响
Table 3 The influence of representation methods in the embedding layer

表示方法	Youtube	Opinion	SMS	E-mail
Random	92.86	86.26	99.64	87.17
SkipGram	93.37	87.39	99.28	86.52
CBOW	93.87	82.50	99.46	87.39
Glove	92.86	87.50	98.56	78.48

实验 3 卷积层激活函数的影响

在 CNN 模型的基础上,本部分实验由卷积层激活函数对于检测准确率的影响进行探究,分别采用 relu 函数、sigmoid 函数、tanh 函数、softplus 函数、softsign 函数与 leakly-relu 函数这 6 种常用的激活函数。

根据表 4 的实验结果,使用不同的激活函数 CNN 模型在不同数据集上表现性能不一。总体而言,使用 sigmoid 函数、tanh 函数与 softsign 函数的 CNN 模型表现性能很差,与每个数据集的最高检测准确率有比较大的差距。而使用 relu 函数、softplus 函数与 leakly-relu 函数的 CNN 模型表现性能较好,其中,relu+CNN 模型在 Opinion 与 SMS 数据集上表现最好,领先其他方法 0.56%与 0.41%;使用 softplus+CNN 模型在 Youtube 与 E-mail 数据集上表现最好,领先其他方法 0.51%与 0.22%。

表 4 卷积层激活函数的影响
Table 4 The influence of activation functions in the convolutional layer

激活函数	Youtube	Opinion	SMS	E-mail
relu	92.86	86.26	99.64	87.17
sigmoid	89.80	75.63	98.75	85.87
tanh	89.28	74.37	99.10	86.96
softplus	93.37	85.63	99.46	87.39
softsign	89.29	74.38	98.56	86.73
leakly-relu	92.36	85.62	99.28	86.95

实验 4 池化层池化策略的影响

在 CNN 模型的基础上,本部分实验对池化

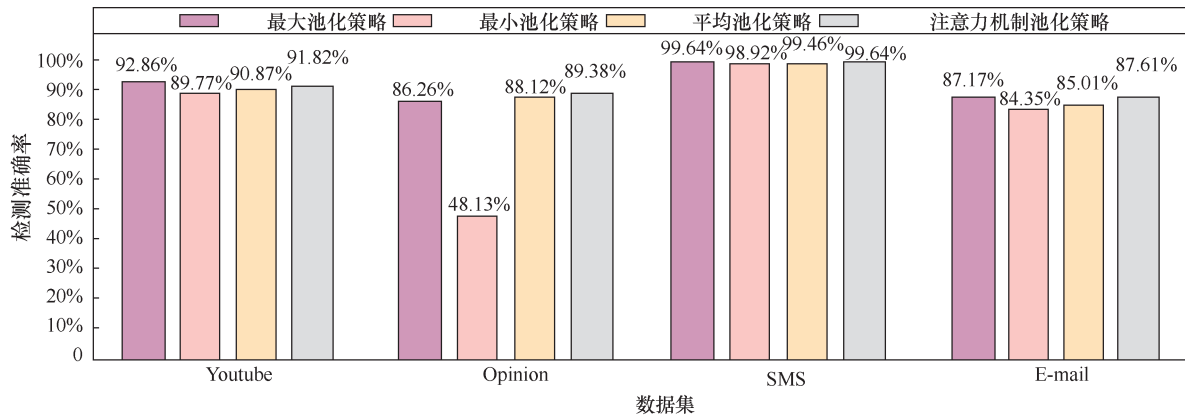


图2 池化层池化策略的影响

Figure 2 The influence of pooling strategies in the pooling layer

层池化策略对检测准确率的影响进行探究, 分别采用最大池化策略、最小池化策略、平均池化策略与注意力机制池化策略这4种池化策略, 如图2所示。

根据式(4)中权值 $a_i \in R^{(k-h_i+1) \times 1}$ 的训练结果取值范围, 注意力机制的池化策略可以模拟出任意的池化策略, 如平均池化策略对应的注意力机制池化策略 a_i 的每个维度的值为 $\frac{1}{k+h_i-1}$ 。理论来说, 基于注意力机制的池化策略的性能好于其他的池化策略。

根据图2的实验结果, 使用不同的池化策略在不同的数据集上表现性能不一。总体而言, 相对于传统的3种池化策略, 基于注意力机制的池化策略在Opinion、SMS、E-mail数据集上取得了最好的效果。但在Youtube数据集上, 相对于最大池化策略, 使用注意力机制策略的CNN的检测准确率下降了1.04%。

4 结束语

针对卷积神经网络方法检测社交网络垃圾信息的噪声性与稀疏性问题, 本文提出了一种注意力机制的卷积神经网络(MA-CNN)检测方法。该方法首先在过滤层融合了关键词提取的思想, 采用了基于朴素贝叶斯权重技术的注意力机制, 降低文本中的噪声含量, 解决了稀疏性问题。其次, 在池化层融合了注意力机制的思想, 采用了基于注意力机制的池化策略, 缓解了稀疏性问题。未来, 将利用

迁移学习的思想进一步解决社交网络垃圾文本检测中的标注瓶颈问题。

参考文献:

- [1] ZAFARANI, REZA, HUAN LIU. 10 bits of surprise: detecting malicious users with minimum information[C]//The 24th ACM International on Conference on Information and Knowledge Management. 2015: 423-431.
- [2] RATKIEWICZ, JACOB, et al. Detecting and tracking political abuse in social media[C]//ICWSM. 2011: 297-304.
- [3] BENEVENUTO, FABRICIO, et al. Detecting spammers on twitter[C]//Collaboration, Electronic Messaging, Anti-abuse and Spam Conference. 2010: 12.
- [4] SONG G E. Short text classification: a survey[J]. Journal of Multimedia, 2014, 9(5): 635.
- [5] YOON K. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
- [6] MIKOLOV, TOMAS, et al. Recurrent neural network based language model[C]//Eleventh Annual Conference of the International Speech Communication Association. 2010.
- [7] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks[C]// International Conference on Machine Learning. 2013: 1310-1318.
- [8] TANG D Y, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification[C]//The 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 1422-1432.
- [9] ZHANG X M, LI H, QU H. AHNN: an attention-based hybrid neural network for sentence modeling[C]//National CCF Conference on Natural Language Processing and Chinese Computing. 2017: 731-740.
- [10] ZHANG Y, ER M J, WANG N. Attention pooling-based convolu-

tional neural network for sentence modelling[J]. Information Sciences 373, 2016: 388-403.

- [11] LI S, ZHAO Z, LIU T, et al. Initializing convolutional filters with semantic features for text classification[C]//The 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 1884-1889.
- [12] DU C S, HUANG L. Text classification research with attention-based recurrent neural networks[C]// International Journal of Computers Communications & Control. 2018: 50-61.
- [13] ZHOU P, SHI W, TIAN J. Attention-based bidirectional long short-term memory networks for relation classification[C]//The 54th Annual Meeting of the Association for Computational Linguistics. 2016: 207-212.
- [14] WANG S D, MANNING C D. Manning. Baselines and bigrams: simple, good sentiment and topic classification[C]//The 50th Annual Meeting of the Association for Computational Linguistics. 2012: 90-94.
- [15] GOLDBERG Y, LEVY O. Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method[J]. arXiv preprint arXiv:1402.3722, 2014.
- [16] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation[C]//The 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[作者简介]



曲强（1994—），男，黑龙江齐齐哈尔人，国家数字交换系统工程技术研究中心硕士生，主要研究方向为网络空间安全、大数据分析与管理、复杂网络异常用户检测。



于洪涛（1970—），男，辽宁丹东人，博士，国家数字交换系统工程技术研究中心研究员，主要研究方向为网络大数据分析与管理。



黄瑞阳（1986—），男，福建漳州人，博士，国家数字交换系统工程技术研究中心助理研究员，主要研究方向为文本挖掘、图挖掘。