



文章编号:1001-9081(2018)11-3063-06

DOI:10.11772/j.issn.1001-9081.2018041356

基于分层注意力机制的神经网络垃圾评论检测模型

刘雨心¹, 王莉^{2*}, 张昊¹

(1. 太原理工大学 信息与计算机学院, 山西 晋中 030600; 2. 太原理工大学 大数据学院, 山西 晋中 030600)

(* 通信作者电子邮箱 591085595@qq.com)

摘要:针对现有垃圾评论识别方法很难揭示用户评论的潜在语义信息这一问题,提出一种基于层次注意力的神经网络检测(HANN)模型。该模型主要由以下两部分组成:Word2Sent 层,在词向量表示的基础上,采用卷积神经网络(CNN)生成连续的句子表示;Sent2Doc 层,基于上一层产生的句子表示,使用注意力池化的神经网络生成文档表示。生成的文档表示直接作为垃圾评论的最终特征,采用 softmax 分类器分类。此模型通过完整地保留评论的位置和强度特征,并从中提取重要的和综合的信息(文档任何位置的历史、未来和局部上下文),挖掘用户评论的潜在语义信息,从而提高垃圾评论检测准确率。实验结果表明,与仅基于神经网络的方法相比,该模型准确率平均提高 5%,分类效果显著改善。

关键词:垃圾评论;表示学习;注意力机制;卷积神经网络;双向长短时记忆

中图分类号:TP183 文献标志码:A

Hierarchical attention-based neural network model for spam review detection

LIU Yuxin¹, WANG Li^{2*}, ZHANG Hao¹

(1. College of Information and Computer, Taiyuan University of Technology, Jinzhong Shanxi 030600, China;

2. College of Data Science, Taiyuan University of Technology, Jinzhong Shanxi 030600, China)

Abstract: Existing measures to detect spam reviews mainly focus on designing features from the perspective of linguistic and psychological clues, which hardly reveal the latent semantic information of the reviews. A Hierarchical Attention-based Neural Network (HANN) model was proposed to mine latent semantic information. The model mainly consisted of the following two layers: the Word2Sent layer, which used a Convolutional Neural Network (CNN) to produce continuous sentence representations on the basis of word embedding, and the Sent2Doc layer, which utilized an attention pooling-based neural network to generate document representations on the basis of sentence representations. The generated document representations were directly employed as features to identify spam reviews. The proposed hierarchical attention mechanism enables our model to preserve position and intensity information completely. Thus, the comprehensive information, history, future, and local context of any position in a document can be extracted. The experimental results show that our method can achieve higher accuracy, compared with neural network-based methods only, the accuracy is increased by 5% on average, and the classification effect is improved significantly.

Key words: spam review; representation learning; attention mechanism; Convolutional Neural Network (CNN); Bidirectional Long-short Term Memory (BLSTM)

0 引言

随着互联网的发展,人们越来越喜欢在网上发表自己的观点,并与其它网络用户分享他们的观点。2016 年,美国 Yelp 评论网站的评论超过 108 万 (<https://www.yelp.com/about>),每年评论数量增加超过 18 万,然而,虚假评论约占 Yelp 总评论的 14%~20%,占 Tripadvisor、Orbitz、Priceline 和 Expedia 总评论的 2%~6%。2011 年美国 Cone Communication 的调查报告 (<http://www.conecomm.com/contentmgr/showdetails.php?id=4008>) 显示,64% 的用户通过阅读相关评论获得产品信息,87% 的用户在阅读肯定评论后购买了此产品,80% 的用户在

阅读否定评论后放弃购买,这充分说明评论对用户的购买决策起到举足轻重的作用,积极的评论可以提高产品口碑和品牌信誉进而提高商家的利润和声誉,垃圾评论在这种背景下应用而生^[1~2]。

垃圾评论是垃圾评论者为了误导潜在客户,精心虚构的虚假评论^[3~4],是商家或用户在个人利益驱使下亲自雇佣水军恶意发布的虚假评论。用户撰写评论的质量受各种因素的影响,如用户的文化背景和用户撰写评论时的情绪。本文垃圾评论不指用户的否定评论,即否定的低质量的评论不一定是垃圾评论。事实上,为了隐藏自己的身份并误导用户,垃圾评论者通常会确保评论的质量,以提高垃圾评论的影响。下

收稿日期:2018-04-30;修回日期:2018-06-26;录用日期:2018-06-29。

基金项目:国家 863 计划项目(2014AA015204);国家自然科学基金资助项目(61702356);山西省自然科学基金资助项目(201703D421013);中国科学院计算技术研究所网络数据科学重点实验室课题(CASNDST20140X)。

作者简介:刘雨心(1984—),女,山西太原人,博士研究生,主要研究方向:数据挖掘、机器学习、深度学习;王莉(1971—),女,山西太原人,教授,博士,主要研究方向:大数据计算与分析、知识图谱、数据挖掘、人工智能;张昊(1988—),男,山西太原人,讲师,博士,主要研究方向:复杂网络。



面是两条来自公开垃圾评论数据集的评论。

1) 如果你在芝加哥,艾尔雷格洛酒店对你来说是完美的。它位于市中心,有时尚的房间和细心的员工。我在酒店住了3个晚上,对一切都很满意。床很舒服,有很多蓬松的枕头,大的平板电视,收音机和iPad 坎站和浴室是干净的。我接触的每个人都非常友好并乐于助人。我在那里的最后一天,我订了房间服务,不仅我的饭菜美味,并按时交付,厨房还打来电话,询问一切是否都好。我从来没有这样的跟进服务。

2) 我在芝加哥希尔顿酒店逗留期间一直很不愉快。你怎么会这样问?好吧,我告诉你,那里的毛巾很脏没有消毒,服务也很糟糕,最糟糕的是,我登记的时候,他们甚至不在桌子上。另外,我从酒店订购了早餐、午餐和晚餐,但我收到的是错误的订单。所有的饭菜,吃完后想吐的感觉。最后,我还为我不想要的东西支付了账单。总的来说,这个酒店对我来说都是非常糟糕和不愉快的。我给它半星的评价。

第1)条不是垃圾评论,即来自顾客的真实的评论;第2)条是垃圾评论,来自土耳其人编写的虚假评论。从上面两条评论可以看出,靠人工从真实的评论中区分垃圾评论是很困难的。在以前的研究中,研究人员邀请三名志愿者识别160条垃圾评论,而志愿者误将垃圾评论判为真实评论,识别准确率仅为53.1%~61.9%^[5],这个结果同样表明垃圾评论不易识别,这导致标注数据不足和难以评价检测结果的困境。因此,垃圾评论检测是一项紧迫必而必要的任务。

用户评论通常是短文本,垃圾评论检测是一个二分类问题,该任务的目标是区分一条评论是否为垃圾评论。现有方法主要遵循文献[6]的工作,采用机器学习的方法来构建分类器,特征工程在这个方向很重要。大部分研究主要集中在从语言学和心理学的角度设计有效的特征以提高分类性能,尽管这些特征表现出强大的性能,但评论的离散型和稀疏性使得研究者们从语篇角度出发,挖掘评论的潜在语义信息变得异常困难。

近年来,在自然语言处理领域,神经网络模型取得了较好成果。基于其良好的性能,一些研究采用神经网络模型来学习文档表示,从而实现从语义的角度检测垃圾评论。例如,Ren等^[7]建立了一个门递归神经网络模型来学习文档表示,虽然取得了较好的效果,但准确率仍有待提高。

基于以上研究,本文提出一种基于层次注意力的神经网络(Hierarchical Attention-based Neural Network, HANN)垃圾评论检测模型,该模型主要由两部分组成:Word2Sent层(见2.1节),在词向量表示的基础上,采用卷积神经网络(Convolutional Neural Network, CNN)^[8]生成连续的句子表示;Sent2Doc层(见2.2节),基于上一层产生的句子表示,使用注意力池化的神经网络生成文档表示,生成的文档表示直接作为垃圾评论的最终特征,采用softmax分类器分类。本文的贡献主要包括以下3个方面:

1) 创新性地提出HANN模型来区分垃圾评论与真实评论,所提模型不需要外部模块,采用端到端的方式进行训练。

2) HANN模型完整地保留了用户评论的位置和强度特征,并从中提取重要的和综合的信息,包括文档中任何位置的历史、未来和局部上下文,从而挖掘用户评论的潜在语义信息。

3) 实验结果表明,与Li等^[9~10]的方法相比,本文方法准确率平均提高5%,在最好的情况下,准确率高达90.9%,比

Li等的方法高出15%,分类效果显著改善。

1 相关工作

与其他类型的垃圾检测,如邮件垃圾^[11]、网页垃圾^[12~13]等相比,由于用户评论具有数量大、噪声多、更新快、主观性高和针对性强等特点,使得用户垃圾评论检测更困难,所以先进的各种垃圾检测方法不能直接用于用户垃圾评论检测。垃圾评论检测被认为是自然语言处理(Natural Language Processing, NLP)领域的一个复杂问题。

2008年,Jindal等^[6]首次提出了垃圾评论这个问题,采用评论内容、评论者和商品本身的特征来训练模型。Jindal等将垃圾评论分为3类,即虚假(负面)评论、仅讨论品牌而非产品的评论以及不存在评论(如广告)的评论,第一类危害性最大也最难识别^[3]。

研究者提出许多垃圾评论检测的方法^[14~15]。大多数研究表明,垃圾评论与真实评论在情感、语言、写作风格、主观性和可读性方面不同^[16~19]。大多数方法在Ott等^[5]最初介绍的合成数据集上进行;但是,文献[20~21]采用相同的方法分别在合成的和真实的数据集上实验,发现合成的数据集是有缺陷的。因为它们没有如实反映真实的垃圾评论,且合成数据集的技术存在问题。

Yoo等^[22]收集了42个虚假的和40个真实的酒店评论,并手动比较了他们的语言差异。Ott等^[23]通过雇佣土耳其人撰写虚假评论来创建数据集,后续研究大都在这个数据集上进行。最近,Liu等^[9]在Ott等工作的基础上发展了一个范围广泛的黄金标准垃圾评论数据集,这个数据集通过众包和领域专家生成,包括3个领域(“酒店”“餐馆”和“医院”),由于此数据集数据量大、覆盖性广,所以本文实验采用这个数据集。

许多方法已经证明,关注评论的上下文相似性是有益的,在这些方法中,重复和近似重复的评论被认为是垃圾评论。Lau等认为垃圾评论者不仅发布虚假评论,而且会以不同的身份复制这些评论作为不同品牌或同一品牌的多种产品的评论,因此,内容相似性是比较研究人员众所周知的技术^[16, 24]。

Heydari等^[25]提出了一个垃圾评论检测系统,评论者的积极性、评价行为和评论的上下文相似性这些特征被综合考虑。从评论的时间序列角度出发,在可疑时间间隔内采用模式识别技术,捕捉垃圾评论;Ahsan等^[26]通过使用评论内容的词频-逆文本频率指数(Term Frequency-Inverse Document Frequency, TF-IDF)特征引入主动学习方法来检测垃圾评论;Zhang等^[27]提出一种基于熵和协同训练算法的CoFea方法,在无标签数据上,采用熵值对所有词汇进行排序,提出两种策略,即CoFea-T和CoFea-S,对比这两种策略后发现CoFea-T策略准确率更高,而CoFea-S策略时间开销少。其他研究也有采用评论内容本身之外的特征,例如,何珑^[28]提出基于随机森林的垃圾评论检测方法,即对样本中的大、小类有放回地重复抽取同样数量样本或者给大、小类总体样本赋予同样的权重以建立随机森林模型,解决只考虑评论特征的选取,忽略了评论数据集不平衡性的问题;Wang等^[29]提出了一种松散的垃圾评论者群体检测技术,该技术采用双向图投影。

以上研究取得了较好的成果,但都表现出一个共同问题:依赖人工设计的、基于特定任务的语言和心理特征,未从文档语篇



的角度有效挖掘用户评论的潜在语义信息。本文提出 HANN 模型,从语篇的角度有效提取文档连续的语义信息,并从中获取重要的和综合的信息,从而提高垃圾评论识别准确率。

2 虚假垃圾评论检测方法

用户评论具有层次结构(单词形成句子,句子形成文档)^[30]。另外,文档中的不同词和句子具有不同的信息量和不同程度的重要性。基于此,本文构建了一个分层注意力神经网络模型来学习文档表示。图 1 描述了模型的结构,主要由两部分组成:Word2Sent 层(见 2.1 节),基于词向量的表示;Sent2Doc 层(见 2.2 节),基于上一层产生的句子表示。生成的文档表示直接作为垃圾评论的最终特征,采用 softmax 分类器分类用户评论。

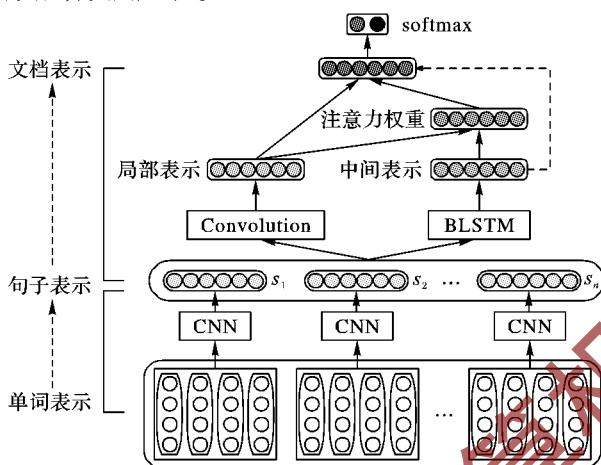


图 1 基于层次注意力机制的神经网络垃圾评论检测模型

Fig. 1 Hierarchical attention-based neural network for spam review detection

2.1 词到句子的表示(Word2Sent layer)

卷积神经网络(CNN)是建模句子语义表示最先进的方法^[31]。CNN 不依赖于外部解析树^[31-32],可用于学习句子的连续表示。卷积操作已被广泛用于合成 N-gram 信息^[33]。N-gram 对许多自然语言处理任务(NLP)有用^[18, 34],本文将 N-gram 应用于 HANN 模型。如图 2 所示,使用 3 个卷积滤波器生成句子表示,因为它们可以捕捉不同粒度的 N-gram 局部语义信息,包括 unigrams、bigrams 和 trigrams。N-gram 在一些 NLP 任务中很强大,比如情感分类^[35]。HANN 模型使用 3 个宽度(width)分别为 2,3 和 4 的卷积滤波器。

正式的定义由 n 个词组成的句子为 $(w_1, w_2, \dots, w_i, \dots, w_n)$ 。每个词 w_i 映射用一个词向量 $e(w_i) \in \mathbb{R}^L$ 表示,卷积滤波器是具有共享参数的线性层列表。 L_1, L_2, L_3 表示 3 个卷积滤波器的宽度。

以 L_1 为例, \mathbf{W}_1 和 \mathbf{b}_1 是该滤波器线性层的共享参数。线性层的输入是在固定长度窗口 L_1 中的词向量表示(word embedding)的连接,表示为 $\mathbf{I}_{1,i} = [e(w_i); e(w_{i+1}); \dots; e(w_{i+L_1-1})] \in \mathbb{R}^{L \times L_1}$ 。

线性层的输出为:

$$\mathbf{H}_{1,i} = \mathbf{W}_1 \cdot \mathbf{I}_{1,i} + \mathbf{b}_1 \quad (1)$$

其中: $\mathbf{W}_1 \in \mathbb{R}^{l_{oc} \times L \times L_1}$, l_{oc} 是线性层的输出大小。将它提供给一个平均池化层,产生一个固定长度的输出向量:

$$\mathbf{H}_1 = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_{1,i} \quad (2)$$

进一步添加一个激活函数 tanh 以合并非线性,滤波器 O_1 的输出如下:

$$O_1 = \tanh(\mathbf{H}_1) \quad (3)$$

类似的,分别得到宽度为 2 和 3 的其他两个卷积滤波器 O_2, O_3 的输出。为了捕捉句子的全局语义信息,用 3 个滤波器的平均输出作为句子的最终输出 S 。

$$S = (O_1 + O_2 + O_3)/3 \quad (4)$$

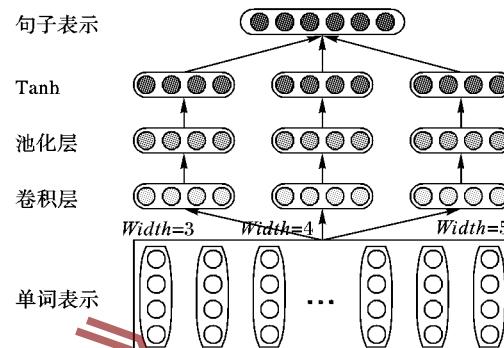


图 2 词到句子的模型

Fig. 2 Word2Sent model

2.2 句子到文档的表示(Sent2Doc layer)

有各种文档表示的方法,如:平均所有的句子表示作为文档的表示,但这种方法不能有效捕捉句子间的语义信息。CNN 采用线性层的共享参数来建模局部句子关系,但 CNN 不能直接对长范围的语篇结构建模,而这对一个文档的表示非常重要。基于上层生成的句子表示,Sent2Doc 层采用注意力池化的 CNN^[36]和双向长短时记忆(Bidirectional Long-Short Term Memory, BLSTM)^[36]模型的组合,实现从语篇的角度提取文档重要的和综合的语义信息。

CNN 是一个功能强大的语义合成模型,卷积操作可以独立地捕获包含在文档中任何位置的信息,但不能捕捉文档长范围的语篇结构,如图 1 所示,卷积滤波器只能对上层产生的文档矩阵执行卷积操作,产生局部表示(Local Representation),再将这个局部表示通过注意力权重(Attention Weight)集成到最终的文档表示中。而注意力权重是通过对局部表示与 BLSTM 生成的中间句子表示(Intermediate Representation)、在训练阶段进行优化而获得的。生成的文档表示作为最终的特征向量输入到顶层 softmax 分类器。在测试阶段,中间句子表示也作为 softmax 分类器的输入,如图 1 中的虚线所示。

在 HANN 模型中,卷积操作是在 k 个滤波器 $\mathbf{W}_c \in \mathbb{R}^{md \times k}$ 和一个连接向量 $\mathbf{x}_{i:i+m-1}$ 之间进行的, $\mathbf{x}_{i:i+m-1}$ 表示从第 i 个句子开始的 m 个句子的窗口。每个滤波器的参数在所有窗口中共享。使用具有不同初始化权重的多个滤波器来提高模型的学习能力。通过交叉验证决定滤波器的数量 k 。卷积运算由 c_i 控制:

$$c_i = g(\mathbf{W}_c^\top \mathbf{x}_{i:i+m-1} + \mathbf{b}_c) \in \mathbb{R}^k \quad (5)$$

其中: $\mathbf{x}_i \in \mathbb{R}^d$, \mathbf{b}_c 是一个偏向量, $g(\cdot)$ 是一个非线性激活函数。本文采用 LeakyReLU^[37] 非线性激活函数,与 ReLU 相比,LeakyReLU 有助于提高学习效率,并且在单元处于非活动状态时允许小的梯度消失。

假定文档的长度为 T ,当句子窗口滑动时,卷积层的特征映射表示如下:

$$\mathbf{c} = [c_1, c_2, \dots, c_T] \in \mathbb{R}^{K \times T} \quad (6)$$



卷积层的输出作为文档的局部表示,每个元素 c_i 都是相应位置的局部表示。

中间文档表示由 BLSTM 生成。BLSTM 是循环神经网络的变体,通过用门控记忆单元代替循环神经网络的隐藏状态,解决 LSTM 的“梯度消失”问题;此外,还可以学习文档任何位置的历史和未来的信息。BLSTM 架构与其他组件一起训练。在训练阶段,损失函数的梯度通过中间文档表示反向传播来优化。

通过对比由卷积操作生成的局部表示与由 BLSTM 生成的中间文档表示来计算注意力权重。为了对比这两种表示,应把局部表示和文档的中间表示映射到同一维空间,本文通过控制 BLSTM 的输出维度与卷积过滤器的数量相同达到这个目的。

文档的中间表示定义为 \bar{d} 。中间表示与每个局部文档表示之间的相似度越高,分配给局部表示的注意力权重越大。注意力权重计算如下:

$$a_i = \frac{\exp(e_i)}{\sum_{i=1}^r \exp(e_i)} \quad (7)$$

其中

$$e_i = \text{sim}(c_i, \bar{d}) \quad (8)$$

术语 a_i 是一个标量,函数 $\text{sim}(\cdot)$ 用于度量两个输入之间的相似度。本文采用余弦相似度。获得注意力权重后,最终的文档表示如下:

$$\bar{d} = \sum_{i=1}^r a_i c_i \in \mathbb{R}^k \quad (9)$$

在识别垃圾评论和真实评论时,评论中的句子在语义表达中扮演着不同的角色,一些句子比另外一些句子更重要。本文中,每个句子的权重代表句子对整个文档含义的贡献,注意力可被视为获得所有句子标注的加权和来计算文档标注。这种方法借鉴了著名的注意力机制思想,将较大的权值赋给较重要的特征,从而提取文档包含的重要信息。

2.3 softmax 分类器

文档表示 \bar{d} 作为顶层分类器的输入。在模型的顶部添加线性转换层将文档表示转换为实值向量 y_c ,softmax 函数将实值向量转换为条件概率,计算如下:

$$P_c = \frac{\exp(y_c)}{\sum_{c' \in c} \exp(y_{c'}')} \quad (10)$$

为了避免过拟合,在模型的倒数第二层,使用掩码概率为 p 的 dropout,dropout 的关键思想是在训练阶段从神经网络中随机丢弃神经单位^[38]。

$$y = \begin{cases} W_s(s \otimes q) + b_s, & \text{training phase} \\ W_s([s, \bar{s}]) + b_s, & \text{testing phase} \end{cases} \quad (11)$$

其中, \otimes 是一个元素乘法运算符; q 是 dropout 率为 p 的掩码向量。在训练阶段实现输出权重 W_s 的 L_2 范数约束。

因为本文模型是监督方法,所以每个文档都有其标签 P_c^g , 使用最小分类交叉熵的目标函数如下:

$$L = - \sum_{i=1}^N \sum_{c=1}^C P_c^g(S_i) \lg (P_c(S_i)) \quad (12)$$

其中: C 是类别数, S_i 表示第 i 个句子。

卷积过滤器、BLSTM 和 softmax 分类器中的所有权重和

偏置都由模型来决定。注意力权重在训练阶段优化。文献 [39] 的 Adadelta 更新规则是一种有效且高效的反向传播算法,本文采用此算法来优化模型。

3 实验结果和分析

在公开的垃圾评论数据集上评价了本文方法的性能,并将该方法与已有方法进行比较,进行了 3 种类型的实验,即领域内、跨领域和混合领域。

3.1 数据集和评价指标

本文采用 Li 等^[9]发布的公开黄金标准垃圾评论数据集,其具体分布见表 1。该数据集包含 3 个领域,即“酒店”“餐馆”和“医生”,每个领域都有 3 种数据类型,分别是“顾客”“专家”和“土耳其人”。真实评论来自具有实际消费体验的“顾客”。垃圾评论由土耳其人和专家编辑,这些专家具有专家级的领域知识。

表 1 三个领域的数据统计

| 领域 | 土耳其人 | 专家 | 顾客 | 领域 | | |
|----|------|-----|-----|------|-----|-----|
| | | | | 土耳其人 | 专家 | 顾客 |
| 酒店 | 800 | 280 | 800 | 医生 | 356 | 0 |
| 餐馆 | 200 | 0 | 200 | | | 200 |

本文采用准确率作为评价指标,所有(顾客/土耳其人/专家)评论都被用于酒店领域中的分类。在餐馆和医生领域中,只有顾客/土耳其人评论被采用,因为专家评论有限。本文使用数据集的 90% 作为训练集,10% 作为测试集。

3.2 Word embedding

本文采用 Word2Vec 工具来表示单词向量。用 skip-gram 和最大化所有词^[40]的平均对数概率的方法,在包括 1000 亿个不同单词的 Google 新闻数据集上训练。每个单词和短语都用 300 维向量表示,词向量矩阵相对较大(3.6 GB),但包含许多不必要的词。具体公式如下:

$$p(\mathbf{y} | \mathbf{x}) = \frac{\exp(\mathbf{x}_i^\top \mathbf{y})}{\sum_{j=1}^v \exp(\mathbf{x}_j^\top \mathbf{y})} \quad (13)$$

其中: c 是上下文窗口大小, T 表示文档的长度。词向量值包含在参数中,在训练过程中优化。

3.3 实验结果分析

3.3.1 领域内结果分析

领域内,根据 Ren 等^[7]的实验设置进行了一组测试并与之对比,顾客/土耳其人/专家评论都用于酒店领域;对于餐馆和医生领域,只有顾客/土耳其人评论被采用,实验结果见表 2。

表 2 两种方法领域内结果

Tab. 2 In-domain results of two methods

| 领域 | 设置 | 方法 | 准确率/% |
|----|------------|---------|-------|
| 酒店 | 顾客/土耳其人/专家 | Ren 等方法 | 80.8 |
| | | 本文方法 | 85.5 |
| 餐馆 | 顾客/土耳其人 | Ren 等方法 | 87.1 |
| | | 本文方法 | 85.0 |
| 医生 | 顾客/土耳其人 | Ren 等方法 | 76.3 |
| | | 本文方法 | 90.9 |



3.3.2 跨领域结果分析

在交叉领域进行两种类型的实验来验证本文模型的泛化能力和领域适应性。在第 1 个实验中, 在一个领域上训练, 分别在另外两个领域测试; 在第 2 个实验中, 在两个领域训练, 在剩下的领域测试。

本文通过在标注丰富的酒店领域数据集上训练模型, 然后分别在餐馆和医生领域测试, 从而评价本文模型的泛化能力和领域适应性。

从表 3 可以看出, Ren 等的方法, 在餐馆领域的测试准确率为 83.5%, 但在医生领域的测试准确率却降到 57.0%。Li 等^[10]方法的准确率在餐馆和医生领域都不太好。本文方法的准确率都优于他们的方法。在餐馆领域, 本文方法获得了最佳结果, 准确率达到了 87.5%; 在医生领域, 准确率最高的是 Li 等^[9]采用离散特征的传统方法。两个先进的神经网络模型的准确率低于 Li 等传统模型的准确率, 而本文模型的准确率与之相近。

表 3 四种方法跨领域结果(在酒店领域训练)

Tab. 3 Cross-domain results of four methods trained by hotel domain

| 领域 | 方法 | 准确率/% | 领域 | 方法 | 准确率/% |
|----|-------------------------|-------|----|-------------------------|-------|
| 餐馆 | Li 等 ^[9] 方法 | 78.5 | 医生 | Li 等 ^[9] 方法 | 74.5 |
| | Ren 等方法 | 83.5 | | Ren 等方法 | 57.0 |
| | Li 等 ^[10] 方法 | 66.8 | | Li 等 ^[10] 方法 | 61.5 |
| | 本文方法 | 87.5 | | 本文方法 | 72.7 |

由于餐馆和酒店之间有许多相似属性, 如环境和位置, 而医生领域与酒店的相似属性少一些, 词汇差异也较大, 这导致在酒店领域训练的模型, 在医生领域的测试结果不如餐馆领域结果。这些结果与以往研究结果一致。

另外, 本文第一次在两个领域上训练, 在剩下的领域测试。例如, 本文在医生和酒店两个领域训练, 在餐馆领域测试。

表 4 显示, 通过使用医生和酒店领域的两组数据进行训练, 在餐馆领域的测试准确率为 77.5%。当只采用酒店领域的数据用于训练时, 在餐馆领域的测试准确率提高了大约 10 个百分点, 因为餐馆领域和酒店领域有许多相似属性, 但与医生领域的相似属性较少, 所以通过在训练过程中添加医生领域的数据, 在餐馆领域的测试准确率不会提高反而降低, 这充分验证了不同的主题在评论中具有不同程度的重要性。例如, 健康信息通常可以成为餐馆评论的强大特征, 因此, 再次验证了本文采用注意力机制方法来挖掘评论中的重要信息是可取的。

表 4 本文方法跨领域结果

Tab. 4 Cross-domain results datasets of proposed method

| 训练领域 | 测试领域 | 准确率/% | 训练领域 | 测试领域 | 准确率/% |
|-------|------|-------|-------|------|-------|
| 医生/餐馆 | 酒店 | 59.5 | 酒店/餐馆 | 医生 | 74.5 |
| 医生/酒店 | 餐馆 | 77.5 | 酒店 | 医生 | 72.7 |
| 酒店 | 餐馆 | 87.5 | | | |

而当采用酒店和餐馆领域的两个数据集训练时, 医生领域的评价准确率为 74.5%, 但是, 如果只采用酒店领域数据训练, 则在医生领域的准确率降低 2%。这表明, 当训练领域的数据集极性与目标评价领域相似度较低时, 使用大量训练数据集可以提高目标领域的评价精度。

3.3.3 混合领域结果分析

在混合领域, 与 Li 等^[10]的方法进行了比较, 其采用来自

土耳其人和专家的所有虚假评论以及顾客的真实评论。同样为了和 Li 等的方法对比, 本文实验设置与他们的方法一致。

Li 等的方法包括段落均值(paragraph-average)、加权平均(weight-average)、句子卷积神经网络(Sentence Convolutional Neural Network, SCNN)、句子加权神经网络(Sentence-Weighted Neural Network, SWNN)以及这些方法和特征的组合。SCNN 是一个基本的文档表示学习模型, 由两个卷积操作组成: 句子卷积通过一个固定长度的窗口为每个句子创建一个组合; 文档卷积把句子向量转换为文档向量。SWNN 是 SCNN 的变体。Li 等采用 KL(Kullback-Leibler) 散度作为一个词的重要性权重来计算一个句子的权重。

本文采用所有句子标注的加权和来计算文档标注。句子的权重衡量句子对整个文档含义的贡献, 评论中的不同句子在文档的语义表示中扮演着不同的角色。从真实的评论中区分垃圾评论时, 一些句子比另一些句子更重要, 因此, 当一个句子对整个文档的含义贡献较大时, 给它分配较大的权重。

表 5 显示本文模型在混合领域取得了最好的结果, 其准确率明显高于其他神经网络。SWNN 模型的准确率为 80.1%, SWNN + 特征 2 的准确率为 82.2%。在垃圾评论检测中, POS(Part-Of-Speech)^[9] 和“第一人称”是强大的特征, 特征 1 指 POS 特征, 特征 2 指 POS + “第一人称”。因此, 可大胆地假设: 如果将这两个特征与本文模型结合, 那么准确率将比对比模型的准确率高出更多。

表 5 各方法混合领域结果

Tab. 5 Mix-domain results of various method

| 模型 | 准确率 | 模型 | 准确率 |
|----------|-------|-------------|-------|
| 段落均值 | 0.729 | SCNN | 0.702 |
| 加权平均 | 0.680 | SWNN | 0.801 |
| 基本长短时记忆 | 0.550 | SWNN + 特征 1 | 0.797 |
| 分层长短时记忆 | 0.618 | SWNN + 特征 2 | 0.822 |
| 基本卷积神经网络 | 0.708 | 本文方法 | 0.855 |

3.3.4 参数分析

在实验中, 本文研究了 3 个参数的影响, 即句子窗口大小、学习率和句子级卷积过滤器的数量。实验结果表明当句子窗口大小设置为 2、3 和 4, 学习率为 0.5, Word2Doc 卷积滤波器数量为 100 时, 准确率最高。

4 结语

一种新的基于分层的注意力机制的神经网络被成功地用于垃圾评论检测。通过使用层次注意力机制, 使评论的位置和强度信息被完整地保留下来。Word2Sent 和 Sent2Doc 的组合使本文模型能从保存的特征中提取重要的和全面的信息, 挖掘用户评论的潜在语义信息, 从而提高垃圾评论识别准确率。本文方法分别在领域内、跨领域和混合领域三个领域上进行了检测对比实验。本文方法准确率比 Li 等^[9-10]的方法准确率平均提高 5%, 最好的情况下, 准确率高达 90.9%, 比 Li 等的方法高出 15%, 总体来说, 本文方法的准确率更高, 泛化能力更强。

将来, 将进一步考虑把从垃圾评论中提取的语言学和心理学特征作为先验知识加入到本文所提出的模型中, 以充分利用两者的优势达到增强分类效果的目的; 另一方面, 可以



将这个新模型扩展到其他 NLP 任务,如情感分析^[4],甚至计算机视觉和图像识别等领域。

参考文献 (References)

- [1] SANTOSH K C, MUKHERJEE A. On the temporal dynamics of opinion spamming: case studies on Yelp[C]// Proceedings of the 25th International Conference on World Wide Web. Montréal, Québec: [s. n.], 2016: 369 – 379.
- [2] 林煜明,王晓玲,朱涛,等.用户评论的质量检测与控制研究综述[J].软件学报,2014,25(3): 506 – 527. (LIN Y M, WANG X L, ZHU T, et al. A review of research on quality inspection and control of user comments [J] . Journal of Software, 2014, 25(3) : 506 – 527.)
- [3] JINDAL N, LIU B. Analyzing and detecting review spam[C]// Proceedings of the 7th IEEE International Conference on Data Mining. Washington, DC: IEEE Computer Society, 2007: 547 – 552.
- [4] 莫倩,杨珂.网络水军识别研究 [J].软件学报,2014, 25(7): 1505 – 1526. (MO Q, YANG K. Overview of Web spammer detection[J] . Journal of Software, 2014, 25(7) : 1505 – 1526.)
- [5] OTT M, CHOI Y, CARDIE C, et al. Finding deceptive opinion spam by any stretch of the imagination[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2011: 309 – 319.
- [6] JINDAL N, LIU B. Opinion spam and analysis[C]// Proceedings of the 2008 International Conference on Web Search and Data Mining. New York: ACM, 2008: 219 – 230.
- [7] REN Y F, JI D H. Neural networks for deceptive opinion spam detection: an empirical study[J]. Information Sciences, 2017, 385: 213 – 224.
- [8] MENG J E, ZHANG Y, WANG N, et al. Attention pooling-based convolutional neural network for sentence modelling[J]. Information Sciences, 2016, 373(C) : 388 – 403.
- [9] LI J, OTT M, CARDIE C, et al. Towards a general rule for identifying deceptive opinion spam [EB/OL]. [2018- 03- 20]. <http://www.aclweb.org/anthology/P/P14/P14-1147.pdf>.
- [10] LI L Y, QIN B, REN W J, et al. Document representation and feature combination for deceptive spam review detection[J]. Neurocomputing 2017, 254: 33 – 41.
- [11] WU Y, FENG G, WANG N, et al. Game of information security investment: impact of attack types and network vulnerability[J]. Expert Systems with Applications, 2015, 42 (15/16): 6132 – 6146.
- [12] FDEZ-GLEZ J, RUANO-ORDAS D, MÉNDEZ J R. A dynamic model for integrating simple Web spam classification techniques [J]. Expert Systems with Applications, 2015, 42 (21): 7969 – 7978.
- [13] GOH K L, SINGH A K. Comprehensive literature review on machine learning structures for Web spam classification[J]. Procedia Computer Science, 2015, 70: 434 – 441.
- [14] JINDAL N, LIU B, LIM E P. Finding unusual review patterns using unexpected rules[C]// Proceedings of the 2010 International Conference on Information and Knowledge Management. New York: ACM, 2010: 1549 – 1552.
- [15] HEYDARI A, TAVAKOLI M, SALIM N. A framework for review spam detection research[EB/OL]. [2018- 03- 20]. <https://pdfs.semanticscholar.org/46a9/74b6a2fe378a366432ac535cf25c9f32d773.pdf>.
- [16] LAU R Y K, LIAO S Y, KWOK C W, et al. Text mining and probabilistic language modeling for online review spam detection [J]. ACM Transactions on Management Information Systems, 2012, 2(4) : 1 – 30.
- [17] PENG Q, ZHONG M. Detecting spam review through sentiment analysis[J]. Journal of Software, 2014, 9(8) : 2065.
- [18] TANG D, WEI F, YANG N, et al. Learning sentiment-specific word embedding for twitter sentiment classification [EB/OL]. [2018- 03- 20]. <http://ir.hit.edu.cn/~dytang/paper/sswe/acl-slides.pdf>.
- [19] 唐晓波,朱娟,杨丰华.基于情感本体和 kNN 算法的在线评论情感分类研究 [J].情报理论与实践,2016 (6): 110 – 114. (TANG X B, ZHU J, YANG F H. Research on online comment emotion classification based on emotion ontology and kNN algorithm [J]. Information Studies: Theory & Application, 2016(6) : 110 – 114.)
- [20] CRAWFORD M, KHOSHGOFTAAR T M, PRUSA J D, et al. Survey of review spam detection using machine learning techniques [J]. Journal of Big Data, 2015, 2(1) : 23.
- [21] ESHRAQI N, JALALI M, MOATTAR M H. Spam detection in social networks: a review[C]// Proceedings of the 2015 2nd International Congress on Technology, Communication and Knowledge. Piscataway, NJ: IEEE, 2015: 148 – 152.
- [22] YOO K H, GRETZEL U. Comparison of deceptive and truthful travel reviews[C]// Proceedings of the 2009 International Conference on Information and Communication Technology. Berlin: Springer, 2009: 37 – 47.
- [23] OTT M, CARDIE C, HANCOCK J T. Negative deceptive opinion spam[EB/OL]. [2018- 03- 20]. <http://www.cs.cornell.edu/Info/People/cardie/papers/NAACL13-Negative.pdf>.
- [24] LIN Y, ZHU T, WU H, et al. Towards online anti-opinion spam: spotting fake reviews from the review sequence[C]// Proceedings of the 2014 International Conference on Advances in Social Networks Analysis and Mining. Washington, DC: IEEE Computer Society, 2014: 261 – 264.
- [25] HEYDARI A, TAVAKOLI M, SALIM N. Detection of fake opinions using time series[J]. Expert Systems with Applications, 2016, 58 (C) : 83 – 92.
- [26] AHSAN M N I, NAHIAN T, KAFI A A, et al. Review spam detection using active learning[C]// Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference. Piscataway, NJ: IEEE, 2016: 1 – 7.
- [27] ZHANG W, BU C Q, YOSHIDA T, et al. CoFea: a novel approach to spam review identification based on entropy and co-training[J]. Entropy, 2016, 18(12) : 429.
- [28] 何珑.基于随机森林的产品垃圾评论识别[J].中文信息学报,2015,29(3): 150 – 154. (HE L. Identification of product review spam by random forest[J]. Journal of Chinese Information Processing, 2015, 29(3) : 150 – 154.)
- [29] WANG Z, HOU T, SONG D, et al. Detecting review spammer groups via bipartite graph projection[J]. Computer Journal, 2016, 59(6): bvx068.
- [30] YANG Z, YANG D, DYER C, et al. Hierarchical attention networks for document classification[EB/OL]. [2018- 03- 20]. <http://www.aclweb.org/anthology/N/N16/N16-1174.pdf>.

(下转第 3074 页)



- 法[J]. 计算机工程与应用, 2014, 50(2): 103 – 106. (ZHANG Z G, JI G L. Parallel algorithm for mining frequent item sets based on FP-Growth [J]. Journal of Computer Engineering and Application, 2014, 50(2): 103 – 106.)
- [9] ZAHNG F, LIU M, GUI F, et al. A distributed frequent itemset mining algorithm using Spark for big data analytics [J]. Cluster Computing, 2015, 18(4): 1493 – 1501.
- [10] DENG L L, LOU Y, et al. Improvement and research of FP-Growth algorithm based on distributed spark[C]// Proceedings of the 2015 International Conference on Cloud Computing and Big Data. Piscataway, NJ: IEEE, 2015: 105 – 108.
- [11] 方向, 张功萱. 基于 Spark 的 PFP-Growth 并行算法优化实现 [J]. 现代电子技术, 2016, 39(8): 9 – 13. (FANG X, ZHANG G X. PFP-Growth parallel algorithm optimization based on Spark [J]. Modern Electronics Technique, 2016, 39(8): 9 – 13.)
- [12] LI C, HUANG X. Research on FP-Growth algorithm for massive telecommunication network alarm data based on Spark[C]// Proceedings of the 2016 IEEE International Conference on Software Engineering and Service Science. Piscataway, NJ: IEEE, 2017: 875 – 879.
- [13] 张稳, 罗可. 一种基于 Spark 框架的并行 FP-Growth 挖掘算法 [J]. 计算机工程与科学, 2017, 33(8): 1403 – 1409. (ZHANG W, LUO K. A parallel FP-Growth mining algorithm based on spark framework[J]. Computer Engineering and Science, 2017, 33 (8): 1403 – 1409.)
- [14] 陆可, 江雨燕, 杜萍萍, 等. 基于 Spark 的并行 FP-Growth 算法优化与实现[J]. 计算机应用与软件, 2017, 34(9): 273 – 277. (LU K, JIANG Y Y, DU P P, et al. Parallel FP-Growth algorithm optimization and implementation based on Spark[J]. Computer Applications and Software, 2017, 34(9): 273 – 277.)
- [15] ZHOU L, WANG X. Research of the FP-Growth algorithm based on cloud environments[J]. Journal of Software, 2014, 9(3): 676 – 682.
- [16] 高权, 万晓东. 基于负载均衡的并行 FP-Growth 算法[J]. 计算机工程, 2018, 39(6): 37 – 72. (GAO Q, WAN X D. Load-balanced parallel FP-Growth algorithm [J]. Computer Engineering, 2018, 39(6): 37 – 72.)
- [17] Frequent itemset mining dataset repository[EB/OL]. [2012-10-21]. <http://fimi.ua.ac.be/data/>.

This work is partially supported by the Science-Technology Program of Hebei Province (17210305D), the Science-Technology Program of Tianjin (16ZXHLSF0023), the Natural Science Foundation of Tianjin (15JCQNJC00600), the Science-Technology Program of Tianjin (15ZXHLLGX00130)

GU Junhua, born in 1966, Ph. D., professor. His research interests include data mining, intelligent information processing, information acquisition and integration, intelligent computing and optimization, function and information display, software engineering, project management.

WU Junyan, born in 1994, M. S. candidate. Her research interests include data mining, computer simulation, machine learning.

XU Xinyun, born in 1995, M. S. candidate. Her research interests include sentiment analysis, natural language processing, deep learning.

XIE Zhijian, born in 1995, M. S. candidate. His research interests include machine learning, data mining.

ZHANG Suqi, born in 1980, Ph. D., lecturer. Her research interests include machine learning, data mining.

(上接第 3068 页)

- [31] KIM Y. Convolutional neural networks for sentence classification [J/OL]. arXiv Preprint, 2014, 2014: arXiv: 1408. 5882 [2014-08-05] [2014-09-03]. <https://arxiv.org/abs/1408.5882>.
- [32] KALCHBRENNER N, GREFENSTETTE E, BLUNSON P. A convolutional neural network for modelling sentences[EB/OL]. [2018-03-20]. <http://mirror.aclweb.org/acl2014/P14-1/pdf/P14-1062.pdf>.
- [33] SANTOS C N D, GATTIT M. Deep convolutional neural networks for sentiment analysis of short texts[EB/OL]. [2018-03-20]. <http://aclweb.org/anthology/C/C14/C14-1008.pdf>.
- [34] REN Y, ZHANG Y, ZHANG M, et al. Improving Twitter sentiment classification using topic-enriched multi-prototype word embeddings[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2016: 3038 – 3044.
- [35] REN Y, ZHANG Y, ZHANG M, et al. Context-sensitive twitter sentiment classification using neural network[C]// Proceedings of the 30th AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2016: 215 – 221.
- [36] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5/6): 602 – 610.
- [37] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models[EB/OL]. [2018-03-20]. http://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf.
- [38] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929 – 1958.
- [39] ZEILER M D. ADADELTA: an adaptive learning rate method [EB/OL]. [2018-03-20]. <http://www.matthewzeiler.com/wp-content/uploads/2017/07/googleTR2012.pdf>.
- [40] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. [2018-03-20]. <http://seed.ucsd.edu/mediawiki/images/e/e3/Wordembeddings.pdf>.

This work is partially supported by the National High Technology Research and Development Program (2014AA015204), the National Natural Science Foundation of China (61702356), the Natural Science Foundation of Shanxi Province (201703D421013), the Key Laboratory Project of Network Data Science and Technology in the Institute of Computing Technology, Chinese Academy of Sciences (CASNDST20140X).

LIU Yuxin, born in 1984, Ph. D. candidate. Her research interests include data mining, machine learning, deep learning.

WANG Li, born in 1971, Ph. D., professor. Her research interests include big data computing and analysis, knowledge mapping, data mining, artificial intelligence.

ZHANG Hao, born in 1988, Ph. D., lecturer. His research interests include complex network.