Journal of Computer Applications
计算机应用, 2020, 40(9): 2536 - 2542
ISSN 1001-9081
CODEN JYIIDU
2020-09-10
http://www.joca.cn

文章编号:1001-9081(2020)09-2536-07
DOI:10.11772/j.issn.1001-9081.2020010048

面向短文本情感分类的端到端对抗变分贝叶斯方法

尹春勇*,章　荪
(南京信息工程大学 计算机与软件学院,南京 210044)
(*通信作者电子邮箱yinchunyong@hotmail.com)

摘　要:针对文本情感分析中文本过短而导致的分类准确度低的问题,结合对抗学习和变分推断提出一种端到端的短文本情感分类模型。首先,使用谱规范化技术解决了判别器在训练过程中的震荡问题;然后,添加额外的分类模型来指导推断模型的更新;其次,使用对抗变分贝叶斯(AVB)模型提取短文本的主题特征;最后,使用三次注意力机制来融合主题特征与预训练词向量特征进行分类。通过在一个产品评论和两个微博数据集上的实验结果证明,所提模型较基于自注意力的双向长短期记忆网络(BiLSTM-SA)在分类准确度上分别提高了2.9、2.2和8.4个百分点。由此可见,该模型适用于挖掘社交短文本中的情感和观点信息,对舆情发现、用户反馈、质量监督和其他相关领域具有重要的意义。

关键词:对抗学习;情感分类;短文本;变分推断;主题模型
中图分类号:TP391.1　　**文献标志码**:A

End-to-end adversarial variational Bayes method for short text sentiment classification

YIN Chunyong*, ZHANG Sun
(*School of Computer and Software*, *Nanjing University of Information Science and Technology*, *Nanjing Jiangsu* 210044, *China*)

Abstract: Concerning the problem of low accuracy in sentiment classification caused by short text, an end-to-end short text sentiment classifier was proposed based on adversarial learning and variational inference. First, the spectrum normalization technology was employed to alleviate the vibration of discriminator in training process. Second, an additional classifier was utilized to guide the updating of the inference model. Third, the Adversarial Variational Bayes (AVB) was used to extract the topic features of the short text. Finally, topic features and pre-trained word vector features were fused by three times of attention mechanism in order to realize the classification. Experimental results on one product review and two micro-blog datasets show that the proposed model improves the accuracy by 2.9, 2.2 and 8.4 percentage points respectively compared to the Bidirectional Long Short-Term Memory network based on Self-Attention (BiLSTM-SA). It can be seen that the proposed model can be applied to mine sentiments and opinions in social short texts, which is significant for public opinion discovery, user feedback, quality supervision and other related fields.

Key words: adversarial learning; sentiment classification; short text; variational inference; topic model

0　引言

近年来随着微博、推特、脸书等社交媒体平台的迅速兴起,互联网已从静态的信息存储知识库转变为充满着不断变化的动态信息论坛[1]。丰富多样的社交媒体平台向用户提供了发表和传播思想与观点的便捷途径,越来越多的用户也逐渐习惯于在线上社交圈中分享个人的日常观点和情感信息。这些社交用户发布的内容可以作为重要的观点和情感信息来源,对舆情发现、用户反馈、观点挖掘、情感分析等领域具有重要的意义[2]。例如,在线上购物网站中,通过对顾客商品评论的挖掘和分析,能够帮助卖家改善服务和商品质量,商品的评论信息也能够帮助其他顾客挑选合适的产品[3]。政府部门同样可以通过收集公众的意见信息,更好地了解公众诉求,并采取积极有效的措施来应对特定的事件。从众多商品评论和社交数据中提取观点和情感信息的过程是复杂的,因此研究者开始关注如何于从海量文本数据中自动提取情感极性和细粒度的类别信息[4-5]。

文本情感分类方法依赖于从文本信息中提取或构造文本的特征表示,目前常用的文本表示方法可以大致分为两类:基于统计的文本表示方法[6]和基于神经网络的文本表示方法[7]。前者通常按照传统的文本处理流程,利用文本的统计信息设计文本特征,这类方法最大的优点在于符合人类语言学的语法规则,易于理解;但是这种方法提取的特征难以消除句子歧义,也会受到数据稀疏问题的影响。后者则是通过训练神经网络,学习以固定维度的连续变量作为词语的分布式表示。这类方法解决了维度爆炸的问题,并且神经网络模型训练得到的中间层参数就可以作为每个词的向量表示。近年来,此类方法在多种文本分析任务上取得了出色的表现,在文本处理领域广受欢迎;但是此类方法依赖于词语的共现性,需要庞大的语料库进行预训练,硬件成本较高。

上述的两种文本表示方法已被广泛应用于句子、段落、文档等其他长文本类型,但是这些方法并不适用于社交媒体中普

收稿日期:2020-01-17;**修回日期**:2020-04-24;**录用日期**:2020-05-06。　**基金项目**:国家自然科学基金资助项目(61772282)。
作者简介:尹春勇(1977—),男,山东潍坊人,教授,博士生导师,博士,主要研究方向:网络空间安全、大数据挖掘、隐私保护、人工智能、新型计算;　章荪(1994—),男,安徽六安人,博士研究生,主要研究方向:机器学习、数据挖掘、文本分类。

遍存在的短文本。与段落或文档不同的是,短文本并不总是遵循自然语言的语法规则,通常会带有多义词和错字,这对正确理解文本语义带来了巨大的挑战。尤其是在许多社交媒体应用场景中,为了提高信息发布和传播的速度,通常会对用户单次发布的文本长度进行限制,例如新浪微博和推特平台限制每条微博或推文长度在140字以内。而在即时通信场景中,平台本身虽然不会对文本长度进行限制,但是大多数用户更倾向于用较短的语句表达自己的想法和观点。由于缺少必要的上下文信息,长文本表示方法难以从短文本中提取正确的语义和情感信息。现有的短文本表示方法可以归纳为三类:1)直接应用长文本表示方法;2)利用概念知识库信息作为补充,消除短文本歧义;3)采用主题建模方法,提取短文本的主题特征表示。

文献[8-9]中处理短文本的方法都是基于外部信息源扩展的思想,作者认为概念化能力是人类所独有的特征,人们能够联想到与文本中词语相关的概念信息。例如,当提到"中国"一词,人们在脑海中会关联到与之相关的概念"国家或地区"。当提到"中国"和"印度"时,则会联想到"亚洲国家"或"发展中国家"。因此通过在概念知识库中查询实体名词的概念信息,能够辅助理解短文本的语义。

但是Chen等[10]指出概念知识扩展的方法具有两个固有的问题。首先概念知识库的信息是有限的,很难确保所有的实体名词都能够查询到对应的概念信息,尤其对于一些罕见的实体名词。另外一点则是当短文本中不含有任何实体名词时,概念化方法则会失效。基于这些考虑,本文没有选择概念化的方法来扩展短文本特征,而是采用主题建模方法。

主题建模方法是以无监督学习的方式对文档中潜在的语义结构进行统计的模型,它通常假设每个文档是由多种主题混合而成,而每一种主题则对应着一种词语的概率分布。经典的主题建模方法隐式狄利克雷模型(Latent Dirichlet Allocation, LDA)正是使用狄利克雷先验来处理文档-主题和单词-主题的分布,从而使得模型具有更好的泛化能力。但是,传统的主题建模方法需要采用变分推断、平均场方法、马尔可夫-蒙特卡洛或吉布斯采样方法来估算后验分布。当这些方法应用在新的主题模型时,即使模型假设只发生细微的改变也需要重新进行完整的推断过程,并且这个过程需要大量的数学推导,限制了主题模型的拓展能力。

Kingma等[11]首次提出了融合自编码器和贝叶斯推断的变分自编码器(Variational Auto-Encoding, VAE),通过训练推断网络直接从文档中学习后验分布的参数,而不需要复杂的数学推导。这个推断网络能够模拟概率推理的过程,也为神经网络模型提供了很好的可解释性。随后Miao等[12]提出了适用于文本处理的神经变分文档模型(Neural Variational Document Model, NVDM),能够提取文档隐含的连续语义特征表示。Srivastava等[13]则继续将神经网络的变分推断功能扩展到经典的LDA模型中。

上述的研究工作都是使用变分自编码器来学习文档的后验分布,而自生成对抗网络(Generative Adversarial Network, GAN)[14]在诸多任务中展现出强大的生成和判别能力后,Mescheder等[15]首次提出了融合GAN和VAE的对抗变分贝叶斯(Adversarial Variational Bayes, AVB)模型,他们认为VAE在图像生成任务中生成模糊图片的原因是推断模型不能正确地捕捉到真实的后验分布。因此,额外的判别器被添加到变分自编码器中来评判学习到的后验概率和真实的后验分布之间的差距。Wang等[16]也进行了相似的改进工作,他们基于GAN的结构,从狄利克雷分布中采集噪声输入到生成器中期望产生真实文档的词频-逆文档频率(Term Frequency-Inverse Document Frequency, TF-IDF)特征表示。

虽然AVB模型通过推断模型和判别器的博弈过程,能够更好地学习到数据的后验分布,但是它的对抗学习过程是基于GAN的思想,同样存在着与GAN相似的问题。GAN在训练过程中经常难以收敛,模型参数会产生剧烈的震荡。Arjovsky等[17-18]指出GAN的问题在于判别器的损失值难以指导训练的进程,需要小心地平衡生成器和判别器的训练程度。此外,在短文本分类任务中,主题模型通常仅作为特征提取器,以文本的词袋(Bag of Words, BoW)表示为输入,向分类模型中输送短文本的主题特征,两个模型的训练过程通常是分开进行的。因此,本文基于AVB模型提出了端到端的短文本情感分类模型——AVBC(Adversarial Variational Bayes Classifier)。首先通过引入谱规范化技术[19]稳定AVB的训练过程,再添加分类模型实现从主题特征提取到分类的端到端过程,并且分类模型的输出同样用于指导推断模型的训练。本文发现传统的主题模型通常仅仅以提取的主题特征作为文本的表示,并没有充分利用预训练词向量的优点。因此在本文提出的AVBC模型中,分类模型以主题特征和预训练词向量作为输入,通过多层的注意力机制融合两种文本表示方法,共同作为情感分类的语义特征。

本文的主要贡献有三点:1)为了解决AVB模型在训练过程中存在的震荡问题,本文引入了谱规范化技术来稳定AVB模型的训练过程;2)本文提出了AVBC模型,实现了从主题建模到情感分类的端到端过程,其中分类模型能够辅助推断网络更好地生成用于下游分类任务的主题表示;3)AVBC模型利用多层的注意力机制融合了主题特征和预训练词向量特征两种文本表示方法,更好地提取短文本的语义特征用于下游情感分类任务。

1　相关工作

正文内容使用神经网络实现对短文本的主题建模是本文的主要工作之一,因此本章将简单回顾VAE和AVB模型相关的背景知识以更好地介绍本文的改进工作。假设样本$\boldsymbol{x}$是由某种随机过程产生的,并且这个生成过程涉及无法被观测的隐变量$\boldsymbol{z}$。这个生成过程包含两步:1)从隐变量先验分布$P(\boldsymbol{z})$中产生一个隐变量$\boldsymbol{z}$;2)从条件分布$P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$中生成数据。隐变量的先验分布通常难以被观测,只能通过已有的样本集合来估测其后验分布,VAE正是利用编码器从样本数据中推断近似的隐变量后验分布$q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$,再利用解码器从近似的隐变量分布中采样生成样本数据。每个样本的边际似然可以表示为:

$$\ln P_{\boldsymbol{\theta}}(\boldsymbol{x}) \geqslant -\mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x}), P(\boldsymbol{z})) + \mathrm{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})} \ln P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \tag{1}$$

式(1)中的右边部分也被称为变分下界或证据下界(Evidence Lower BOund, ELBO),它能保证样本的边际似然概率始终大于ELBO,因此在使用最大似然求解参数$\boldsymbol{\varphi}$和$\boldsymbol{\theta}$时,可以通过最大化变分下界来实现:

$$\max_{\boldsymbol{\theta}} \max_{\boldsymbol{\varphi}} \mathrm{E}_{P_D(\boldsymbol{x})}[-\mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x}), P(\boldsymbol{z})) + \mathrm{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})} \ln P_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] \tag{2}$$

其中P_D是样本的分布。通常变分下界的质量依赖于推断模型近似的后验分布$q_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$,在VAE和NVDM模型中都是假设其是一个满足对角方差矩阵的正态分布,它的平均值$\boldsymbol{\mu}$和方差$\boldsymbol{\sigma}^2$都可以通过向推断模型中输入样本得到,如图1所示。

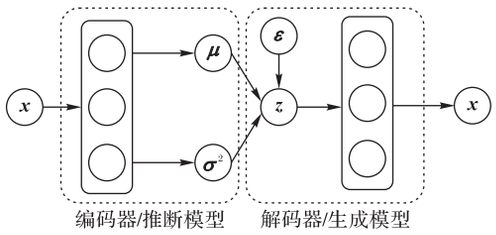

图1 VAE模型结构示例

Fig. 1 Structure of VAE model

VAE模型首先使用编码器从输入样本中学习隐变量的后验分布参数平均值$\boldsymbol{\mu}$和方差$\boldsymbol{\sigma}^2$,再从该分布中采样得到隐变量$\boldsymbol{z}$。由于采样操作是不可导的,因此文献[12]中提出了重参数化技巧,将从正态分布$\boldsymbol{z}\sim\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\sigma}^2)$中采样的过程替换为从标准正态分布中采集噪声$\boldsymbol{\varepsilon}$,再利用参数变换$\boldsymbol{z}=\boldsymbol{\mu}+\boldsymbol{\varepsilon}\times\boldsymbol{\sigma}$得到隐变量。重参数化的变换使得采样过程可导,进而使得模型得以训练。AVB模型则是在VAE的基础上将目标函数优化重写为:

$$\max_{\theta}\max_{\varphi}\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}[\ln P(\boldsymbol{z})-\ln q_\varphi(\boldsymbol{z}|\boldsymbol{x})+\ln P_\theta(\boldsymbol{x}|\boldsymbol{z})] \tag{3}$$

AVB的核心思想在于训练一个判别器$T(\boldsymbol{x},\boldsymbol{z})$来度量先验分布$P(\boldsymbol{z})$和后验分布之间$q_\varphi(\boldsymbol{z}|\boldsymbol{x})$的距离,因此确定后验分布后,判别器的训练目标即是:

$$\max_{T}\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}\ln\sigma(T(\boldsymbol{x},\boldsymbol{z}))+\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{P(\boldsymbol{z})}\ln(1-\sigma(T(\boldsymbol{x},\boldsymbol{z}))) \tag{4}$$

其中,σ表示逻辑回归(sigmoid)激活函数。通过训练,分类模型T能够对来自于目标分布$P_D(\boldsymbol{x})P(\boldsymbol{z})$的数值对$(\boldsymbol{x},\boldsymbol{z})$给出较小的实数值;相反的,对于来自于推断模型的组合$P_D(\boldsymbol{x})q_\varphi(\boldsymbol{z}|\boldsymbol{x})$返回较大的数值。AVB模型中推断模型和判别器的对抗训练是基于GAN模型的思想,因此当固定后验分布和条件分布时,判别器的最优解是:

$$T^*(\boldsymbol{x},\boldsymbol{z})=\ln q_\varphi(\boldsymbol{z}|\boldsymbol{x})-\ln P(\boldsymbol{z}) \tag{5}$$

将判别器的最优解代入到原目标函数式(3)中,可以得到推断模型和生成模型的目标函数为:

$$\max_{\theta}\max_{\varphi}\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}[-T^*(\boldsymbol{x},\boldsymbol{z})+\ln P_\theta(\boldsymbol{x}|\boldsymbol{z})] \tag{6}$$

因此,通过式(4)和(6)不断迭代训练判别器、推断模型和生成模型,AVB模型最终将达到纳什均衡。VAE和AVB的目的很相似,都是希望构建一个从隐变量分布到目标数据的生成模型,它们都是在进行概率分布之间的转换。为了训练这个模型,首先需要度量这两种分布之间的距离。从VAE和AVB的目标函数可以看到,VAE使用KL散度(Kullback-Leibler Divergence)来直接度量估计的后验分布和先验分布之间的距离,而AVB则是利用神经网络训练一个判别器来度量二者之间的距离,当AVB的判别器达到最优时,将近似于KL散度。

AVB模型通过引入GAN的对抗训练思想解决了VAE无法捕捉真实后验分布的问题,但是对抗训练的过程中判别器的梯度会产生剧烈的震荡,给模型的训练带来了不稳定性。在对抗模型中,生成器的更新依赖于从判别器向后传递的梯度,收敛过快的判别器传递的梯度将很小,使得生成器不能有效更新,而收敛过慢的判别器则不能正确地指导生成器的优化方向。

此外,在文本处理领域,VAE和AVB这些无监督的模型虽然可以提取文本的主题特征,但是通常是作为特征提取器为下游的分类任务输送特征,这个过程需要分布进行。基于主题建模的短文本分类方法通常以词频(Term Frequency,TF)、BoW、或TF-ID表示的文本统计特征作为输入,忽略了预训练词向量的作用,因此本文提出使用谱规范化技术稳定AVB中判别器的训练,引入分类模型以推断模型提取的主题特征和预训练词向量特征作为输入,利用多级注意力对二者进行融合,同时分类模型也会参与指导推断模型和生成模型的训练,最终实现更精准的短文本情感分类。

2 对抗变分贝叶斯分类

本文提出的AVBC模型整体结构如图2所示,该模型由推断模型、生成模型、评分模型和分类模型这4个子模型组成。推断模型采用AVB中方法,以d_x维原始数据$\boldsymbol{x}$和服从标准正态分布的d_ε维噪声$\boldsymbol{\varepsilon}$作为输入,学习隐变量的后验分布$q_\varphi(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\varepsilon})$,输出推断得到的$d_z$维隐变量$\boldsymbol{z}$。在文本处理领域,主题模型通常以BoW作为文本的特征表示,而本文使用TF作为文本表示,因为推断模型是从文档-词的概率分布中学习文档-主题分布,使用词频能够更好地表示文档-词特征,但是本文并未选择TF-IDF,因为它不符合文档之间相互独立的假设。关于BoW、TF、TF-IDF三种统计文本表示对模型的影响,将在后续的实验部分展开定量的分析和对比。

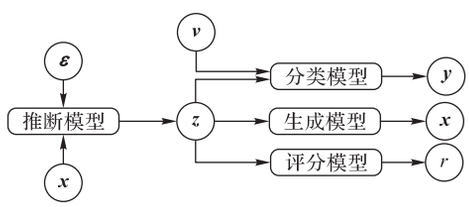

图2 AVBC模型框架

Fig. 2 Framework of AVBC model

生成器则以隐变量作为输入,最大化似然概率$P_\theta(\boldsymbol{x}|\boldsymbol{z})$,期望生成原始数据。本文取消判别器的最后一个非线性激活层,使得判别器转变为能够学习后验分布和先验分布距离$r\in[-\infty,+\infty]$的评分器。为了在提取文本主题特征的同时实现对短文本的情感分类,AVBC模型中添加了一个以双向长短期记忆(Bi-directional Long Short-Term Memory, BiLSTM)网络为基础构成的分类模型,它以主题特征和d_v维词向量特征$\boldsymbol{v}$为输入,输出文本对应的预测情感类别$\boldsymbol{y}$。

2.1 谱规范化

首先为了解决AVB模型在训练过程中判别器震荡的问题,判别器取消了最后一个非线性激活层,期望作为评分器输出后验分布与先验分布的瓦瑟斯坦(Wasserstein)距离,而非GAN中对于输入真假数据的判断。此时评分器的目标函数式(4)将改变为:

$$\max_{\|T\|_{\mathrm{Lip}}\leqslant k}\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{q_\varphi(\boldsymbol{z}|\boldsymbol{x})}T(\boldsymbol{x},\boldsymbol{z})-\mathrm{E}_{P_D(\boldsymbol{x})}\mathrm{E}_{P(\boldsymbol{z})}T(\boldsymbol{x},\boldsymbol{z}) \tag{7}$$

GAN中判别器不稳定的问题在文献[17]中得到了充分的解释和讨论,给出的解决方法是对判别器权重进行约束,使得其满足k-利普希茨(k-Lipschitz)约束条件:

$$\|T\|_{\mathrm{Lip}}=\frac{\|T(a)-T(b)\|}{\|a-b\|}\leqslant k \tag{8}$$

通过式(8)的约束能够保证判别器函数梯度的变化速率始终控制在常数值k的范围内,从而确保判别器在更新过程中的稳定性。因此,如何使得判别器满足利普希茨约束是实现判别器稳定更新的关键。文献[17]中给出的方法是通过对判别器权重的修剪实现的,梯度修剪是在判别器每一次更新

后,将所有的神经网络参数修剪到固定区间,通常是[-1,1],这种方法虽然简单,但是会使得大部分的参数被局限在-1或1,导致判别器对生成器的变化不再敏感,进而引起模式崩塌问题。文献[18]中则在此基础上提出不使用强制性的参数修剪,而是在判别器损失函数上添加对其梯度的惩罚项进行约束。本文选用的则是文献[19]中提出的谱规范化技术,通过约束AVBC中评分器参数的谱范数来实现利普希茨约束。对于给定的矩阵$\boldsymbol{A}$,其谱范数为:

$$\sigma(\boldsymbol{A}) = \max_{h\neq 0}\frac{\|\boldsymbol{A}\mathrm{h}\|}{\|h\|} = \max_{h\leqslant 1}\|\boldsymbol{A}h\| \tag{9}$$

从式(9)可以看到,矩阵的谱范数约束了矩阵操作的变化范围,因此函数$f(x)=\boldsymbol{A}x/\sigma(\boldsymbol{A})$是满足于1-Lipschitz约束条件的。对于仅由全连接层$\boldsymbol{W}^i$和线性整流(Rectified Linear Unit, ReLU)函数R^i的评分器T,R^i可以看作是局部线性的,且已经满足1-Lipschitz约束条件,所以只要对每一个全连接层进行谱规范化处理,则能够保证评分器整体满足1-Lipschitz约束条件,方法如下式所示:

$$\overline{\boldsymbol{W}}^i_{\mathrm{SN}} = \frac{\boldsymbol{W}^i}{\sigma(\boldsymbol{W}^i)} \tag{10}$$

2.2　三次注意力融合

在传统的文本主题分类模型中,主题模型只作为特征提取器,经过预先的训练后提取主题特征,向下游的分类模型中提供输入数据。此外上游的主题模型和下游的分类模型通常任务目标不同,二者独立存在,因此本文提出在AVB基础上添加分类模型与推断模型相连,推断模型为分类模型提供主题特征输入,分类模型的预测结果将辅助推断模型产生具有区分性的主题特征,二者共同训练,实现了从主题推断到文本分类的端到端学习。

本文的分类模型结构设计如图3所示,该分类器以推断模型提取的隐变量即主题特征$\boldsymbol{z}$和预训练模型提取的$d_v\times t$文本词向量特征$\boldsymbol{v}=\{v_1,v_2,\cdots,v_t\}$作为输入,其中$t$为文本词语数量,$d_v$为词向量嵌入特征维度。为了融合主题特征和词向量特征,本文在不同阶段使用了三次注意力机制进行融合。

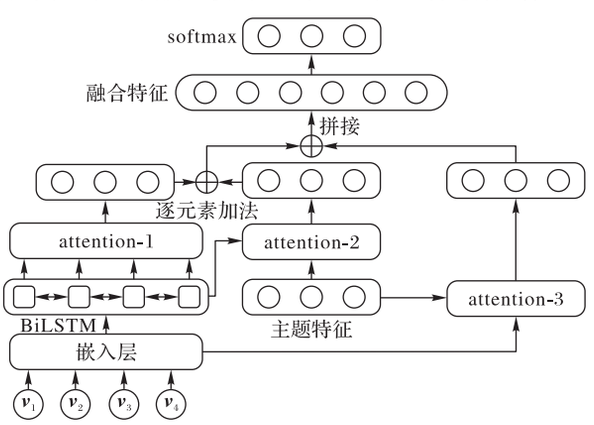

图3　AVBC中分类模型结构

Fig. 3　Structure of classifier in AVBC

文本的词向量特征首先被输入到隐藏单元数量为d_h的BiLSTM中,提取前向和后向隐藏状态提取的$2d_h\times t$时序特征$\boldsymbol{h}=[\boldsymbol{h}_{1:t};\boldsymbol{h}_{t:1}]$。第一次的注意力融合使用vanilla注意力[20],通过计算主题特征与每个$2d_h$维隐藏状态特征$\boldsymbol{h}_i$的相似性,对隐藏状态特征赋予权重α_i,也即是根据推断模型提取的主题特征计算每个隐藏状态的重要程度:

$$a_i = \sigma(\boldsymbol{W}_2\tanh(\boldsymbol{W}_1[\boldsymbol{h}_i;\ \boldsymbol{z}])) \tag{11}$$

其中,形状为$d_a\times(2d_h+d_z)$的矩阵$\boldsymbol{W}_1$和d_a维向量$\boldsymbol{W}_2$都是不带偏置的全连接层权重矩阵,d_a为超参数。本文在3次计算注意力权重过程中并没有按照传统的方法,使用softmax激活函数对所有的权重进行归一化处理,而是仿照压缩激励模型(Squeeze-and-Excitation Network, SENet)[21]中利用sigmoid函数直接计算权重。因此在计算每个隐藏状态特征的t维权重向量$\boldsymbol{attn}_1=[\alpha_1;\alpha_2;\cdots;\alpha_t]$后,通过矩阵乘法得到第1次注意力融合后的$2d_h$维特征$\boldsymbol{s}_1$。第1次注意力融合的特征利用主题特征作为查询向量,更加关注于与主题相关的上下文特征。

$$\boldsymbol{s}_1 = \boldsymbol{attn}_1\boldsymbol{h}^{\mathrm{T}} = \sum_{i=1}^{t}\alpha_i\boldsymbol{h}_i \tag{12}$$

第2次的注意力权重计算则是利用隐藏状态特征计算自注意力,即利用文本特征的上下文时序特征计算重要程度。自注意力机制被广泛应用于自动翻译领域[22],本文选择的自注意力计算方法则是使用Lin等[23]提出的source2token自注意力机制。如下式所示:

$$\beta_i = \sigma(\boldsymbol{W}_4\tanh(\boldsymbol{W}_3\boldsymbol{h}_i)) \tag{13}$$

相似的,$d_a\times 2d_h$维矩阵$\boldsymbol{W}_3$和d_a维$\boldsymbol{W}_4$都是不带偏置的全连接层权重矩阵,通过计算可得到第2次注意力t维权重$\boldsymbol{attn}_2=[\beta_1;\beta_2;\cdots;\beta_t]$,同样经过矩阵乘法运算得到融合后的$2d_h$维特征$\boldsymbol{s}_2=\boldsymbol{attn}_2\boldsymbol{h}^{\mathrm{T}}$。第2次注意力融合后的特征利用文本特征自身的时序特征,消除无意义词汇的干扰,更加关注于与语义相关的上下文特征。

第3次的注意力融合同样是基于vanilla注意力,利用主题特征作为查询向量,以词向量特征作为键值向量,计算方式如下所示:

$$\gamma_i = \sigma(\boldsymbol{W}_6\tanh(\boldsymbol{W}_5[\boldsymbol{v}_i;\ \boldsymbol{z}])) \tag{14}$$

利用第3次注意力权重集合$\boldsymbol{attn}_3$对词向量特征进行加权求和,得到融合后的d_v维特征$\boldsymbol{s}_3=\boldsymbol{attn}_3\boldsymbol{h}^{\mathrm{T}}$。第3次的注意力融合是利用主题特征直接对词向量进行选择,计算每个词语与主题的相关性。

在获得3次注意力融合后的语义特征后,需要进一步融合这3种特征用于预测类别目标。本文通过实验选择的融合方式如下所示:

$$\boldsymbol{s} = [\boldsymbol{s}_1 + \boldsymbol{s}_2;\ \boldsymbol{W}_7\boldsymbol{s}_3 + \boldsymbol{b}_7] \tag{15}$$

首先,第1次和第2次注意力融合的特征经过逐元素加法进行相加,因为它们都是从隐藏状态特征上提取的,分别表示着与主题特征和上下文特征的重要性,相加后能够综合表示每一个时刻隐藏状态的重要程度。$\boldsymbol{s}_3$经过线性映射放缩后和$\boldsymbol{s}_1+\boldsymbol{s}_2$经过拼接操作融合成最终的$4d_h$维语义特征$\boldsymbol{s}$。最后,融合了从主题到上下文、从主题到词向量特征以及上下文自注意力的语义特征被输入到全连接网络中,以softmax激活函数输出对情感类别的预测$P(\boldsymbol{y}|\boldsymbol{s})$,如下式所示:

$$P(\boldsymbol{y}|\boldsymbol{s}) = \mathrm{softmax}(\boldsymbol{W}_9\tanh(\boldsymbol{W}_8\boldsymbol{s} + \boldsymbol{b}_8) + \boldsymbol{b}_9) \tag{16}$$

2.3　端到端学习

在本文提出的AVBC模型中,推断模型、生成模型、评分模型和分类模型这4个模型互相协作,共同训练以实现从主题建模到情感分类的端到端模型,它的实现依赖于协调每个模型的训练过程。额外添加的分类器除了实现融合主题特征和词向量特征外,还用于指导推断模型的更新方向,期望主题特征能够更加适合于分类任务。因此本文修改了推断模型的原目标函数式(6)为:

$$\max_{\theta}\max_{\varphi}\mathrm{E}_{P_D(x)}\mathrm{E}_{q_\varphi(z|x)}[-T(\boldsymbol{x},z)+\ln P_\theta(\boldsymbol{x}|z)+\lambda P(\boldsymbol{y})\ln P_\omega(\boldsymbol{y}|z,\boldsymbol{v})] \tag{17}$$

其中：$P_\omega(\boldsymbol{y}|z,\boldsymbol{v})$为分类模型的预测输出，$\boldsymbol{\omega}$表示分类模型的参数，使用交叉熵$P(\boldsymbol{y})\ln P_\omega(\boldsymbol{y}|z,\boldsymbol{v})$作为分类模型的损失函数，同时分类模型的结果也用于更新推断模型的参数，使用惩罚项系数λ来调节推断模型目标函数的效果。

为了实现不同模型的共同训练，本文通过赋予不同的学习率来调节每个模型的训练进程。调节模型的训练进行方法还可以通过多步训练的方式，尤其是在GAN的对抗训练中，判别器的性能直接影响到生成器的效果。然而判别器的收敛通常比生成器慢，所以在文献[19]中，作者提出在迭代训练时每更新判别器k次再训练一次生成器。多次训练判别器可以通过赋予较大的学习率来实现，并且本文发现AVBC模型中分类模型的收敛速度过快，需要使用更小的学习率。因此在训练过程中固定评分模型的学习率为lr，推断模型和生成模型的学习率为$0.1\times lr$，分类模型的学习率则为$0.01\times lr$，四个模型都使用RMSProp(Root Mean Square Prop)优化器。AVBC模型的训练步骤如下所示：

步骤1 从样本数据、先验分布和标准正态分布中分别采样n次构成样本集合、隐变量集合与噪声集合；

步骤2 利用式(17)计算推断模型的损失值，利用式(6)计算生成器损失值，按照式(4)计算判别器损失值，最后利用交叉熵计算分类器损失值；

步骤3 依照模型各部分损失值，计算梯度并利用优化器更新模型参数；

步骤4 重复步骤1~3直到模型收敛。

3 实验

在本章中，将通过实验来展示改进模型的效果和性能，从定性和定量两方面对其展开分析。所有实验代码使用Python3.6编写，深度学习框架选用PyTorch1.3.0，机器学习方法使用Scikit-learn0.21.3提供的API。实验平台为Ubuntu 18.04操作系统，使用Intel Core i9-9900K CPU@3.6 GHz ×16处理器和GeForce RTX 2080 GPU用于加速模型训练。

本文的研究工作主要针对于短文本的情感分类任务，为了检验模型的分类效果，选用3个采集自真实应用环境中的数据集。

NLPCC2013 来自于2013年国际自然语言处理和中文计算会议(Natural Language Processing and Chinese Computing, NLPCC)的中文微博情感分析任务，用于从中文微博短文本数据中识别情感的细粒度类别，包括愤怒、厌恶、悲伤、恐惧、惊讶、喜欢和高兴这7个类别。

NLPCC2014 来自于2014年NLPCC会议的中文微博文本情感分析任务，同样是从微博短文本中识别细粒度的情感类别，共有7类。

Product Review 来自于2014年NLPCC会议的深度学习情感分类任务，用于从商品评论数据中提取消极和积极两种情感极性信息。

实验参考文献[10]的设置对原始数据进行预处理，并且所有的数据集按照80%、10%、10%的比例划分训练、验证和测试集。本文选用HanLP(Han Language Processing)作为分词工具，与文献[10]中使用搜狗语料库训练词向量模型不同的是，本文使用Li等[24]提供的预训练模型。考虑到本文的测试数据集多是来自于微博平台，因此选用在微博语料库上预训练的300维词向量模型。经过预处理后的数据集信息如表1所示，前两个数据集统一每条短文本词语数量为20，最后一个数据集设置词语数量为40。

表1 数据集统计信息

Tab. 1 Statistics of three datasets

数据集	类别数	样本数	词汇数	平均词数
NLPCC2013	7	4937	14605	17
NLPCC2014	7	2017	7475	17
Product Review	2	10000	30396	40

3.1 主题特征分类效果测试

为了检验AVBC模型提取的主题特征在分类任务上的表现，本文利用3种常用的分类器：支持向量机(Support Vector Machine, SVM)、K近邻(K-Nearest Neighbor, KNN)、决策树(Decision Tree, DT)的分类准确度作为评判指标。直接调用SciKit-Learn模块实现这3种算法，所有参数使用默认设置。用于对比的主题模型如下所示。

LDA 经典的主题模型，基于狄利克雷先验分布，利用词语在文档中的共现性产生主题的词分布；

VAE 首个结合了自编码器和变分推断的神经网络模型，利用神经网络学习后验分布的参数，再结合重参数化技巧采集隐变量生成原始数据；

ProdLDA(Products Latent Dirichlet Allocation) 在VAE的基础上将先验分布的假设由多元正态分布替换为对数正态分布，更加符合LDA中的先验假设，代码的实现来自于GitHub平台分享；

AVB 首次提出使用对抗博弈的方式来训练变分自编码器，训练判别器识别隐变量先验分布和推断模型后验分布的差距，指导推断模型和生成模型的更新，代码实现同样来自于GitHub平台分享。

除以上4个对比模型外，本文还在相同设置下使用无监督训练的AVBC-无监督作为对比，它只使用谱规范化等技术优化评分模型，分类模型不参与指导推断模型的更新。所有基于神经网络的主题模型统一学习率为0.01，训练次数为50，隐变量数量为10，在Product Review数据集上的实验对比结果如表2所示。

表2 主题特征分类效果对比

Tab. 2 Comparison of topic feature classification effect

模型	BoW			TF			TF-IDF		
	SVM	KNN	DT	SVM	KNN	DT	SVM	KNN	DT
LDA	0.641	0.595	0.562	0.510	0.512	0.499	0.513	0.520	0.483
VAE	0.512	0.504	0.502	0.515	0.522	0.505	0.511	0.536	0.530
ProdLDA	0.551	0.508	0.495	0.500	0.518	0.490	0.497	0.517	0.507
AVB	0.554	0.535	0.513	0.531	0.527	0.507	0.510	0.503	0.516
AVBC-无监督	0.587	0.536	0.546	0.632	0.579	0.576	0.592	0.557	0.550
AVBC	0.732	0.735	0.724	0.778	0.759	0.745	0.726	0.723	0.721

从实验结果可以看到,没有分类模型参与推断模型训练的AVBC在多数情况下能够取得较好的分类结果,但是使用BoW表示文本特征的LDA模型在SVM分类器上取得了无监督主题模型中最好的结果。当分类模型参与指导推断模型的更新时,提取的主题特征在3种分类器上的结果均获得了大幅的提升。这个实验不仅对比了不同模型提取的主题特征在分类任务上的表现,还比较了使用BoW、TF和TF-IDF作为主题模型输入的效果。可以看到,使用TF表示的文本特征在本文提出的AVBC模型上能够取得最好的结果,因此在后续实验对比中使用TF特征作为输入。

3.2　注意力可视化展示

在本文提出的AVBC模型中,分类模型以主题特征和词向量特征共同作为输入,利用三次不同阶段的注意力融合两种输入特征,提取最终的语义特征用于分类。本文并未采用softmax激活函数归一化注意力权重,而是直接使用sigmoid激活函数将注意力权重映射到[0,1]区间内,即允许注意力权重的累加和超过1。为了更好地可视化注意力权重,从AVBC中输出的注意力权重将使用极大极小归一化(Min Max Scaler)进行缩放,如图4所示。

上下两个热力图分别表示Product Review数据集中消极和积极样本的示例,热力图的每一列表示样本中的一个词语,第1、2、4行分别对应三次计算获得的注意力权重系数α、β和γ,第3行则表示前两次注意力逐元素相加的结果。从图4中可以看到,三次注意力权重大多聚焦在与目标情感类别相关的词语上,这也证明了本文注意力计算方法的有效性。

	制冷	效果	超差	,	而且	噪音	太大	!	有点	上当受骗	的	感觉
α	1	0.64	0.51	0.58	0.72	0.66	0.45	0.29	0.15	0.24	0	0.18
β	1	0.6	0.49	0.44	0.55	0.60	0.40	0.32	0.30	0.39	0	0.47
$\alpha+\beta$	1	0.6	0.49	0.45	0.55	0.60	0.40	0.31	0.30	0.38	0	0.45
γ	0.84	0.038	0.88	0.49	0.04	1	0.82	0.38	0.52	1	0	0.39

(a) 消极样本

	曲目	很	有	代表性	,	非常	不错	,	值得	收藏
α	0.94	0	0.43	0.42	0.34	0.18	0.13	0.33	0.36	1
β	1	0.13	0.34	0.10	0.15	0.10	0.10	0	0.07	0.61
$\alpha+\beta$	1	0	0.33	0.14	0.15	0.05	0.03	0.04	0.10	0.74
γ	1	0.24	0.16	0.63	0	0.17	0.67	0	0.81	0.82

(b) 积极样本

图4　注意力权重可视化

Fig. 4　Visualization of attention weights

3.3　情感分类效果测试

本文主要的研究目标是实现对短文本的情感分类,所以本文利用3个情感相关的数据集进行验证。对比方法选用的是在文本分类领域常用的几种深度学习方法。

TextCNN(Text Convolution Neural Network):Kim[25]首次提出了使用卷积神经网络在预训练词向量特征上提取文本特征,是文本分类领域常用的基准方法;

BiLSTM:利用双向的LSTM结构学习文本前向和后向的时序特征,能够更好地从文本中提取上下文关系;

BiLSTM-MP(BiLSTM-Max Pooling):Lee等[26]提出的这个模型是用于短文本的序列分类任务,在BiLSTM的基础上学习文本的双向时序特征,使用最大池化层在所有的隐藏状态上提取句子的特征,再利用多个全连接层实现短文本的分类;

BiLSTM-SA(BiLSTM-Self Attention):使用最大池化层提取句子特征的方法会忽略掉重要的细节信息,因此该模型对所有的隐藏状态计算注意力权重,利用加权累加后的结果作为句子特征。

TextCNN方法的学习率固定为0.01,训练次数为50,另外3种以BiLSTM为基础的模型则固定学习率为1×10^{-4},在Product Review、NLPCC2013、NLPCC2014这3个数据集上的训练次数依次为50,100和200。本文提出的AVBC模型固定学习率为0.01,训练次数依次为50,50和100,式(17)中的惩罚项超参数λ分别设置为1、0.15和0.1,固定主题数量为10,噪声维度为4,实验对比结果如表3所示。

从表3的实验结果可以发现,本文提出的AVBC模型在3个数据集上都取得了最好的分类准确度,这证明了本文算法在短文本情感分类任务上的有效性。本文实验中所有方法的效果均优于文献[10]中的实验结果,甚至包括几种常用的基准方法,这主要的原因是因为该文献中使用的是搜狗语料库训练的50维词向量特征,而本文的词向量选用的是微博语料库训练的300维词向量特征,这也说明了融合预训练词向量特征对短文本分类的重要性。

表3　分类准确度对比

Tab. 3　Comparison of classification accuracy

模型	Product Review	NLPCC2013	NLPCC2014
TextCNN	0.738	0.466	0.520
BiLSTM	0.754	0.449	0.515
BiLSTM-MP	0.758	0.482	0.525
BiLSTM-SA	0.752	0.464	0.510
AVBC	0.781	0.486	0.594

4　结语

为了解决短文本情感分类任务中存在的数据稀疏和信息匮乏的问题,本文提出了基于对抗变分贝叶斯模型的AVBC模型,利用谱规范化技术解决判别器在训练过程中的震荡问题,引入分类模型实现从主题建模到情感分类的端到端学习过程,并利用分类结果指导推断模型的更新以提取更有助于下游分类任务的主题特征。此外,AVBC的分类模型以主题特征和预训练词向量特征共同作为输入,利用三次不同阶段的注意力融合输入特征,提取最终的文档表示用于情感分类。在三个真实环境采集的社交文本数据集上的验证效果可以证明,本文提出的AVBC模型在短文本情感任务上能够取得出色的表现。

然而,AVBC模型中各部分模型的共同训练依赖于自定义的学习率设置,需要平衡各部分模型的训练速度。分类模型的引入虽然证明了具有指导推断模型训练的作用,但是惩罚项系数的选择影响着模型的整体训练效果。在未来的工作中,将尝试研究如何以自适应的方法动态调整模型的训练过程,更好地实现各部分模型的协作和共同训练。此外,在单一

文本模态的信息不足以准确预测情感类别时,可以考虑融合其他模态的信息进行补充。因此,多模态情感分析工作将作为后续研究的工作重点。

参考文献(References)

[1] 刘德喜,聂建云,万常选,等. 基于分类的微博新情感词抽取方法和特征分析[J]. 计算机学报, 2018 ,41(7):1574-1597. (LIU D X, NIE J Y, WANG C X, et al. A classification based sentiment words extracting method from microblogs and its feature engineering [J]. Chinese Journal of Computers, 2018, 41(7): 1574-1597.)

[2] GIACHANOU A, CRESTANI F. Like it or not: a survey of twitter sentiment analysis methods [J]. ACM Computing Surveys, 2016, 49(2): No. 28.

[3] YADOLLAAHI A, SHAHRAKI A G, ZAIANE O R. Current state of text sentiment analysis from opinion to emotion mining [J]. ACM Computing Surveys, 2017, 50(2): No. 25.

[4] 曾义夫,蓝天,吴祖峰,等. 基于双记忆注意力的方面级别情感分类模型[J]. 计算机学报, 2019,42(8):1845-1857. (ZENG Y F, LAN T, WU Z F, et al. Bi-memory based attention model for aspect level sentiment classification [J]. Chinese Journal of Computers, 2019, 42(8): 1845-1857.)

[5] 许银洁,孙春华,刘业政. 考虑用户特征的主题情感联合模型[J]. 计算机应用, 2018, 38(5):1261-1266. (XU Y J, SUN C H, LIU Y Z. Joint sentiment/topic model integrating user characteristics [J]. Journal of Computer Applications, 2018, 38(5): 1261-1266.)

[6] PASSALIS N, TEFAS A. Learning bag-of-embedded-words representations for textual information retrieval [J]. Pattern Recognition, 2018, 81: 254-267.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171-4186.

[8] WANG J, WANG Z, ZHANG D, et al. Combining knowledge with deep convolutional neural networks for short text classification [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2017: 2915-2921.

[9] JIA C, CARSIN M B, WANG X, et al. Concept decompositions for short text clustering by identifying word communities [J]. Pattern Recognition, 2018, 76: 691-703.

[10] CHEN J, HU Y, LIU J, et al. Deep short text classification with knowledge powered attention [C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI, 2019: 6252-6259.

[11] KINGMA D P, WELLING M. Auto-encoding variational Bayes [EB/OL]. [2019-12-20]. https://arxiv. org/pdf/1312. 6114. pdf.

[12] MIAO Y, YU L, BLUNSOM P. Neural variational inference for text processing [C]// Proceedings of the 33rd International Conference on Machine Learning. New York: JMLR. org, 2016: 1727-1736.

[13] SRIVASTAVA A, SUTTON C. Autoencoding variational inference for topic models [EB/OL]. [2019-03-04]. https://arxiv. org/pdf/1703. 01488. pdf.

[14] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[15] MESCHEDER L, NOWOZIN S, GEIGER A. Adversarial variational Bayes: unifying variational autoencoders and generative adversarial networks [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 2391-2400.

[16] WANG R, ZHOU D, HE Y. ATM: adversarial-neural topic model [J]. Information Processing and Management, 2019, 56(6): No. 102098.

[17] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 214-223.

[18] GULRAJANI I, AHEMD F, ARJOVSKY M, et al. Improved training of Wasserstein GANs [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2017: 5767-5777.

[19] MIYATO T, KATAOKA T, KOYAMA M, et al. Spectral normalization for generative adversarial networks [EB/OL]. [2020-02-16]. https://arxiv. org/pdf/1802. 05957. pdf.

[20] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [EB/OL]. [2019-05-19]. https://arxiv. org/pdf/1409. 0473. pdf.

[21] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. , 2017: 6000-6010.

[23] LIN Z, FENG M, DOS SANTOS C N, et al. A structured self-attentive sentence embedding [EB/OL]. [2019-03-09]. https://arxiv. org/pdf/1703. 03130. pdf.

[24] LI S, ZHAO Z, HU R, et al. Analogical reasoning on Chinese morphological and semantic relations [C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2018: 138-143.

[25] KIM Y. Convolutional neural networks for sentence classification [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2014: 1746-1751.

[26] LEE J Y, DERNONCOURT F. Sequential short-text classification with recurrent and convolutional neural networks [C]// Proceedings of the 2016 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2016: 515-520.

This work is partially supported by the National Natural Science Foundation of China (61772282).

YIN Chunyong, born in 1977, Ph. D., professor. His research interests include cyberspace security, big data mining, privacy protection, artificial intelligence, new computing.

ZHANG Sun, born in 1994, Ph. D. candidate. His research interests include machine learning, data mining, text classification.