

基于 word2vec 和双向 LSTM 的情感分类深度模型*

黄贤英¹, 刘广峰^{1†}, 刘小洋^{1,2}, 阳安志¹

(1. 重庆理工大学 计算机科学与工程学院, 重庆 400054; 2. College of Engineering, University of Alabama, Tuscaloosa, Alabama 35401, USA)

摘要: 针对社交网络文本传统情感分类模型存在先验知识依赖以及语义理解不足的问题,提出一种基于 word2vec 和双向长短时记忆循环神经网络的情感分类模型——WEEF-BILSTM。采用基于 CBOW (continuous bag-of-words) 方式的 word2vec 模型针对语料训练词向量,减小词向量间的稀疏度,通过双向 LSTM 神经网络获取更为完整的文本上下文信息从而提取出深度词向量特征,继而使用 one-versus-one SVM 对其进行情感分类。实验结果表明,提出的 WEEF-BILSTM 模型较其他模型分类效果更好,能达到更优的准确率和 F 值。

关键词: 文本分类; 情感分析; 双向长短时记忆循环神经网络; 词向量; 社交网络

中图分类号: TP391 **文献标志码:** A **文章编号:** 1001-3695(2019)12-013-3583-05

doi:10.19734/j.issn.1001-3695.2018.08.0599

Sentiment classification depth model based on word2vec and bi-directional LSTM

Huang Xianying¹, Liu Guangfeng^{1†}, Liu Xiaoyang^{1,2}, Yang Anzhi¹

(1. College of Computer Science & Engineering, Chongqing University of Technology, Chongqing 400054, China; 2. College of Engineering, University of Alabama, Tuscaloosa Alabama 35401, USA)

Abstract: Aiming at the problem of prior knowledge and lack of semantic understanding in the traditional sentiment classification model of social network text, this paper proposed a sentiment classification model called WEEF-BILSTM based on word2vec and bi-directional long short-term memory recurrent neural network. It used the CBOW-based word2vec model for corpus to train word vectors, reduced the sparsity between word vectors, and obtained more complete text context information through the bi-directional LSTM neural network to extract the depth word vector features, and then used one-versus-one SVM to classify emotions. The experimental results show that the proposed WEEF-BILSTM model has better classification results than other models and can achieve better accuracy and F value.

Key words: text classification; sentiment analysis; Bi-LSTM; word embedding; social network

互联网带动了社交平台和电商网站的兴起,从而使得人们可以自由地对自己感兴趣的事物发表观点,这些观点都含有用户所表达的情感信息。分析这些蕴涵情感的观点不仅有利于在舆情监控、营销策略等方面的应用,还可以用于问答系统的个性化定制。精准地进行情感分析已成为当前工业界和学术界的迫切要求,这也大力促进了情感分析成为自然语言处理领域的一个热点研究方向。情感分析又称意见挖掘、观点分析等,是通过计算机帮助用户快速获取、整理互联网上海量的主观评价信息,对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程^[1]。作为情感分析的一项核心工作,情感分类主要是针对文本所蕴涵的情感(如积极、消极、中性等)进行自动分类。目前,国内外学者主要基于传统的机器学习和现在流行的深度学习中的浅层神经网络两种方法进行情感分类。张向阳等人^[2]基于情感词典和复杂网络人工构建情感分类特征,继而利用传统的支持向量机、朴素贝叶斯等机器学习算法进行情感分类;Li 等人^[3]通过人工构建句子的标签和上下文标签特征进行情感分类。这类基于传统机器学习的方法需要大量人工标注数据且过度依赖背景知识。陈翠平^[4]通过深度信念网络来对文本提取情感特征进而完成情感分类任务;Kim^[5]基于卷积神经网络进行情感分类,取得了比较好的效

果。这类基于深度学习的方法效果都优于人工构建特征的方法,但在捕捉深层语义信息方面的能力仍然有所欠缺。

综上所述,传统情感分类模型需要大量的先验知识作为背景且在语义理解方面存在信息缺失,而词向量表示方法能够在更抽象的层面上描述原始输入数据的特征分布情况,双向 LSTM 神经网络可以更完整地捕捉文本上下文信息,故基于 word2vec 和双向 LSTM,本文提出一种情感分类模型 WEEF-BILSTM (sentiment classification model based on word embedding extension feature and bi-directional LSTM)。

1 相关工作

1.1 深度学习与 LSTM

Hinton 等人于 2006 年首次提出深度学习的概念,认为深度学习中模拟人脑的信号处理机制可以用于复杂数据的特征提取,且提取的特征可以在更高层面上抽象地表示原始数据,对于解决比较抽象的识别类任务有很大帮助。近年来深度学习领域蓬勃发展,已在计算机视觉和自然语言处理等多个领域取得许多优秀成果。Li 等人^[6]基于循环神经网络解决了句子中评价对象的情感极性判别问题;Kalchbrenner 等人^[7]基于 DCNN (dynamic convolutional neural network) 模型来识别句子中

收稿日期: 2018-08-21; 修回日期: 2018-09-29 基金项目: 重庆理工大学研究生创新项目(yx2018245); 重庆市教育委员会人文社会科学研究项目(17SKG144); 国家教育部人文社科青年基金资助项目(16YJC860010); 国家社科基金资助项目(17XXW004); 2018 年重庆市科委技术创新与应用示范项目(cstc2018jcsx-msybX0049)

作者简介: 黄贤英(1967-),女,教授,硕士,主要研究方向为自然语言处理、网络信息传播; 刘广峰(1995-),男(通信作者),硕士研究生,主要研究方向为自然语言处理、机器学习、深度学习(liuguangfeng@heysky.com); 刘小洋(1980-),男,副教授,硕士,博士(后),主要研究方向为社交网络、信息传播、计算机应用等; 阳安志(1993-),男,硕士研究生,主要研究方向为网络信息传播、机器学习、深度学习。

的情感极性; Dos Santos 等人^[8]基于 CNN 构建字符表示层从而达到了文本分类的目的; Krizhevsky 等人^[9]同样基于卷积神经网络设计出用于 ImageNet 挑战赛的神经网络结构,大幅度提升了分类性能。本文涉及的 LSTM 网络是目前应用最为广泛的深度学习结构之一,是对来源于人工神经网络的循环神经网络(recurrent neural network, RNN)的一个扩展。在自然语言处理领域,上下文关系的处理极其重要,而 RNN 虽然适用于上下文处理,但其内部由于梯度消失/爆炸导致不能对长期依赖进行针对性处理。为了弥补 RNN 对于长期依赖的不足, LSTM 网络应运而生,其自然行为便是长期的保存输入。本文尝试利用 LSTM 神经网络解决自然语言处理领域中的情感分类问题。

1.2 LSTM 与情感分类

随着人工智能浪潮袭来,作为人工智能领域中一个重要方向,自然语言处理越来越受到学术界和工业界的重视。不同于图像和语音,文本在很多方面有其自身的特点。例如短文本本身的稀疏性和不规范性使得自然语言处理在研究过程中遇到了更多的困难。然而随着深度学习的高速发展,国内外研究者已将深度学习与自然语言处理领域紧密结合起来。本文所涉及的情感分类问题正是其中一大研究热点。

Zhu 等人^[10]基于 LSTM 对评论语句建模成词序列从而解决了情感分类问题。Ren 等人^[11]利用 LSTM 同时结合主题特征对 Twitter 短文本进行特征提取,在实验中也取得了很大进步。但纯粹基于 LSTM 解决情感分类问题过程中同样存在梯度消失和爆炸等问题,在上下文的记忆关联上也容易丢失信息,而双向 LSTM 的出现正好解决了这个问题。Chen 等人^[12]基于 CRF 算法进行分词继而通过双向 LSTM 和 CNN 获得词向量表示从而达到情感分类的目的。任勉等人^[13]基于 CBOW 模型训练词向量,然后将词向量作为双向 LSTM 网络的输入对神经网络进行训练从而进行情感分类。然而,如何准确科学地设计双向 LSTM 神经网络,提高情感分类准确性,是使用双向 LSTM 进行情感分类的关键难点,也是本文研究的切入点和突破口。

1.3 word2vec 模型

word2vec 模型是在 Log-Bilinear 和 NNLM 两个模型的基础上由 Mikolov 等人^[14]开发的工具,可以将词从高维空间分布式地映射到低维空间且保留了词向量之间的位置关系,从而解决了向量稀疏和语义联系两个问题,其分为 CBOW (continuous bag-of-words) 和 skip-gram 两种方式。CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量,而输出就是这个特征词对应的词向量。skip-gram 模型和 CBOW 的思路正好相反,即输入是一个特征词对应的词向量,而输出是这个特征词对应的上下文词向量。本文采用 CBOW 模型来训练词向量。

在 CBOW 模型中,给定词 k 对应的上下文向量 $\text{context}(k)$, 根据 $\text{context}(k)$ 去预测 k 。若指定 $\text{context}(k)$ 是一个正样本,其余词作为负样本,于是可以通过负采样得到关于 k 的负样本集 $\text{Neg}(k)$ 。词的特征可表示为

$$F^k(k) = \begin{cases} 1 & k = k \\ 0 & k \neq k \end{cases} \quad (1)$$

CBOW 模型目标函数可表示为

$$J(k) = \prod_{c \in (k) \cup \text{Neg}(k)} P(c | \text{context}(k)) \quad (2)$$

$$\text{其中: } P(c | \text{context}(k)) = \begin{cases} \sigma(X_k^T \theta^c) & F^k(c) = 1 \\ 1 - \sigma(X_k^T \theta^c) & F^k(c) = 0 \end{cases} \quad (3)$$

2 提出的研究方法

本文提出一种基于 word2vec 和双向 LSTM 的情感分类深度模型。首先预先对语料库中的文本进行预处理,将其转换为规范化文本;然后使用 word2vec 获取文本的词向量表示,并将

其作为本文设计的神经网络模型的输入,通过训练网络模型获取对应的深度词向量特征;最后将深度词向量特征使用 one-versus-one SVM 进行模型的训练和分类。

2.1 文本预处理

由于评论文本与传统媒体发布的正式规范的文本不一样,这种文本通常很短且最多不会超过 130 字,包含标点、俚语、URL、特定术语缩写、用户昵称等内容,这些内容给情感分类带来了很大的噪声干扰。

为了去除不必要的噪声干扰,本文使用自然语言处理领域的相关技术对数据集中的文本进行预处理。首先基于 jieba 分词工具对每一个评论文本进行分词;然后基于哈尔滨工业大学以及百度提供的停用词表对其进行去停用词处理,继而进行去除噪声处理。在去除噪声时,本文主要针对其中涉及到的俚语、特定术语缩写、用户昵称、URL、标点符号等字符串进行处理。

2.2 文本词向量表示

本文采用谷歌开源工具 word2vec 基于 CBOW 模型对语料进行训练从而获得文本词向量表示。词向量可捕捉从语料中的词语到实数维向量空间的复杂映射,指定词向量空间为 Ψ , 其大小为 $|\Psi| \times m$, Ψ 中每一行表示某个单词的 m 维词向量, $|\Psi|$ 表示词向量中包含词语的个数。语料库中的一条评论文本 T 可表示为如下序列:

$$(t_1, t_2, \dots, t_n) \quad (4)$$

其中: n 表示文本 T 中的词语个数; t_i 表示 T 中第 i 个词语 ($1 \leq i \leq n$)。若将 T 转换为词向量矩阵,首先应在 Ψ 中搜索词 t_i 对应的词向量,若存在则选中对应的词向量,用 Z_i 表示,否则将对应的词向量即 Z_i 置为 0。在找到每一个词语对应的词向量之后,将每一个词向量堆叠形成词向量特征矩阵 Z ,其大小为 $n \times m$, Z 的每一行表示语料库中一个词语对应的词向量,可表示为

$$(t_1, t_2, \dots, t_n) \Rightarrow (Z_1, Z_2, \dots, Z_n)^T \quad (5)$$

2.3 WEEF-BILSTM 情感分类模型

如图 1 所示,本文构建的 WEEF-BILSTM 模型由输入层、词嵌入层、双向 LSTM 层、聚合层、最大池化层、全连接层、分类层组成,每一层的输出为下一层的输入。利用该模型即可获取情感分类的深度词向量特征,模型构造方法如下:

a) 输入层。该层为模型的输入部分,即将语料库中的一段文本 T 输入进行后续处理。

此处假定有一篇文档的文本 T 为“新包装感觉还不错”。在经过数据预处理之后,输入文本可表示为

$$T = [\text{'新'}, \text{'包装'}, \text{'感觉'}, \text{'还'}, \text{'不错'}] \quad (6)$$

b) 词嵌入层。通过 word2vec 对语料库的训练得到上下文向量列表,然后在上下文向量列表中搜索输入文本的各个词语所对应的词向量并将其合并。

依此方式可将输入序列 T 表示为

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1m} \\ Z_{21} & Z_{22} & \dots & Z_{2m} \\ Z_{31} & Z_{32} & \dots & Z_{3m} \\ Z_{41} & Z_{42} & \dots & Z_{4m} \\ Z_{51} & Z_{52} & \dots & Z_{5m} \end{pmatrix} \quad (7)$$

其中: Z 中的第 i 行代表输入文本 T 中的第 i 个词语所对应的 m 维词向量。

c) 双向 LSTM 层。该层相当于特征抽取部分,通过构造两个 LSTM 神经网络来实现从两个相反的方向获取信息,更有利于从整体上捕捉句子的长依赖关系以及文本的深层语义表达,两个神经网络的输入一致。LSTM 的优势在于其具有三种特

殊的门函数:输入门、遗忘门、输出门,通过这三种门来控制神经网络的记忆。单个 LSTM 记忆单元(如该层的 LSTM_{*j*})在某一时刻 t 的前向计算过程为^[15]

遗忘门机制:

$$f_t = \sigma(W_f \times [h_{t-1} \ x_t] + b_f) \quad (8)$$

输入门机制:

$$i_t = \sigma(W_i \times [h_{t-1} \ x_t] + b_i) \quad (9)$$

$$\tilde{C}_t = \tanh(W_C \times [h_{t-1} \ x_t] + b_C) \quad C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (10)$$

输出门机制:

$$o_t = \sigma(W_o \times [h_{t-1} \ x_t] + b_o) \quad Z_t = o_t \times \tanh(C_t) \quad (11)$$

其中: $\{W, b\}$ 是神经网络训练的参数集合; $\tilde{C}_t, f_t, i_t, o_t$ 分别表示时刻 t 记忆单元的输入单元、遗忘门、输入门和输出门的输出值; h_{t-1}, x_t 分别表示时刻 t 上一个记忆单元以及当前记忆单元的输入; C_t 表示时刻 t 记忆单元的内部状态; Z_t 表示时刻 t 记忆单元的输出。

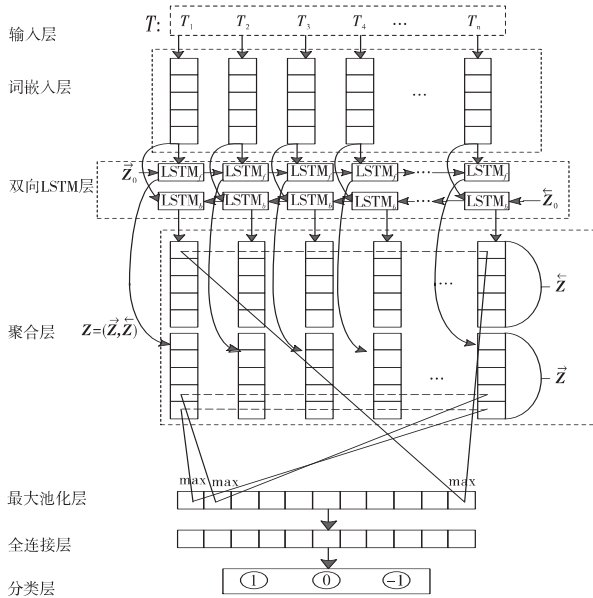


图1 WEFF-BiLSTM 情感分类模型

基于上述结构,该层操作为

$$\tilde{Z}_t = \text{LSTM}_f(x_t, \tilde{Z}_{t-1}) \quad \tilde{Z}_t = \text{LSTM}_b(x_t, \tilde{Z}_{t+1}) \quad (12)$$

其中: LSTM_{*f*}、LSTM_{*b*} 分别表示 LSTM 神经网络的前向传播以及反向传播; \tilde{Z}_t, \tilde{Z}_t 分别表示前向 LSTM 以及反向 LSTM 在时刻 t 的输出向量。

经过双向 LSTM 层处理之后,词嵌入之后的 Z 就变成了如下形式:

$$\tilde{Z} = \begin{pmatrix} \tilde{Z}_{11} & \tilde{Z}_{12} & \cdots & \tilde{Z}_{1c} \\ \tilde{Z}_{21} & \tilde{Z}_{22} & \cdots & \tilde{Z}_{2c} \\ \tilde{Z}_{31} & \tilde{Z}_{32} & \cdots & \tilde{Z}_{3c} \\ \tilde{Z}_{41} & \tilde{Z}_{42} & \cdots & \tilde{Z}_{4c} \\ \tilde{Z}_{51} & \tilde{Z}_{52} & \cdots & \tilde{Z}_{5c} \end{pmatrix} \quad \tilde{Z} = \begin{pmatrix} \tilde{Z}_{11} & \tilde{Z}_{12} & \cdots & \tilde{Z}_{1c} \\ \tilde{Z}_{21} & \tilde{Z}_{22} & \cdots & \tilde{Z}_{2c} \\ \tilde{Z}_{31} & \tilde{Z}_{32} & \cdots & \tilde{Z}_{3c} \\ \tilde{Z}_{41} & \tilde{Z}_{42} & \cdots & \tilde{Z}_{4c} \\ \tilde{Z}_{51} & \tilde{Z}_{52} & \cdots & \tilde{Z}_{5c} \end{pmatrix} \quad (13)$$

此处的列数 c 代表 LSTM 单元的神经元数目。

d) 聚合层。该层主要是将上一层得到的前向传播输出向量以及反向传播输出向量拼接起来,即

$$Z_t = \text{concat}(\tilde{Z}_t, \tilde{Z}_t) \quad (14)$$

经过聚合层处理之后,上一层的 \tilde{Z}_t, \tilde{Z}_t 整合为如下形式:

$$Z = \begin{pmatrix} \tilde{Z}_{11} & \tilde{Z}_{12} & \cdots & \tilde{Z}_{1c} & \tilde{Z}_{11} & \tilde{Z}_{12} & \cdots & \tilde{Z}_{1c} \\ \tilde{Z}_{21} & \tilde{Z}_{22} & \cdots & \tilde{Z}_{2c} & \tilde{Z}_{21} & \tilde{Z}_{22} & \cdots & \tilde{Z}_{2c} \\ \tilde{Z}_{31} & \tilde{Z}_{32} & \cdots & \tilde{Z}_{3c} & \tilde{Z}_{31} & \tilde{Z}_{32} & \cdots & \tilde{Z}_{3c} \\ \tilde{Z}_{41} & \tilde{Z}_{42} & \cdots & \tilde{Z}_{4c} & \tilde{Z}_{41} & \tilde{Z}_{42} & \cdots & \tilde{Z}_{4c} \\ \tilde{Z}_{51} & \tilde{Z}_{52} & \cdots & \tilde{Z}_{5c} & \tilde{Z}_{51} & \tilde{Z}_{52} & \cdots & \tilde{Z}_{5c} \end{pmatrix} \quad (15)$$

e) 最大池化层。该层主要进行最大池化操作以获取向量中最显著的特征值,在一定程度上降低了数据稀疏性对分类器性能的影响。同时由于每条输入文本包含的词语数量不一致,通过池化操作也可得到定长的特征向量 M_t 。计算方式如下所示:

$$M_t = \max\{Z_t(i)\} \quad 1 \leq i \leq c \quad (16)$$

具体操作如图 2 所示。图 2 中左面的矩形框表示经过聚合层处理得到的矩阵向量,此处选择池化单元的宽度及其高度均为 2,步长同样设置为 2,经过池化处理后,原矩阵空间就变成了右面的矩形框所示的矩阵空间。

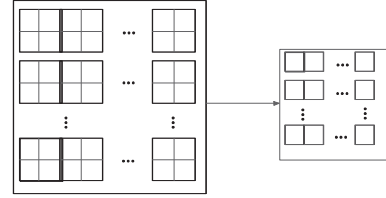


图2 最大池化示意图

至此就完成了一篇文档文本数据的特征提取工作。

f) 全连接层。上述过程描述了双向 LSTM 的特征提取过程,而所有文档的特征会在全连接层汇聚为最终用于情感分类的深度词向量特征 M :

$$M = \begin{pmatrix} M_1 \\ M_2 \\ M_3 \\ M_4 \\ \vdots \\ M_n \end{pmatrix} \quad (17)$$

其中: $M_i (1 \leq i \leq n)$ 表示第 i 篇文档对应的深度词向量特征。

g) 分类层。在训练神经网络时,分类层采用 softmax 函数。全连接层输出的特征 M 会利用 softmax 输出情感极性的判别类别(正面 1、中性 0、负面 -1),且在训练过程中采用反向传播算法对网络中的参数进行梯度更新。

2.4 情感分类模型的训练和判别

当整个 WEFF-BiLSTM 神经网络模型梯度更新收敛后,即可获得深度词向量特征(即全连接层输出的特征)用做情感分类特征,具体处理算法如图 3 所示。进而使用 SVM 对训练集和测试集中的样本对应的深度词向量特征进行模型训练和分类判别。

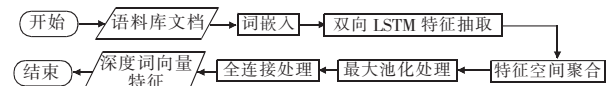


图3 深度词向量特征处理算法

SVM 是一个分类算法,通过找到一个分类平面,将数据分隔在平面两侧,基于结构风险最小化原理从而达到分类的目的。SVM 是一种具有很好泛化能力的预测工具,现已经被广泛应用于人脸识别、文本分类等领域。然而本文所进行的情感分类任务面向的是一个三分类问题,传统的 SVM 算法不适用,因此采用 one-versus-one SVM 进行分类,在任意两个类别之间设计一个 SVM 分类器,总计需要设计三个 SVM 分类器。当对一个未知样本 testcase 进行情感极性判别时,使用设计的三个 SVM 分类器进行分类判决投票,得票数最多的类别即为该样本的情感类别。具体的分类判决投票算法如下:

Pos、Neu、Neg 分别为积极、中性、消极三个类别的训练样本,经训练后生成三个分类器 classifier₁、classifier₂、classifier₃。

$$\text{classifier}_1 = \text{SVM}(\text{Pos}, \text{Neu}) \quad (18)$$

$$\text{classifier}_2 = \text{SVM}(\text{Pos}, \text{Neg}) \quad (19)$$

$$\text{classifier}_3 = \text{SVM}(\text{Neu}, \text{Neg}) \quad (20)$$

初始化 $\text{positive} = \text{neutral} = \text{negative} = 0$,根据下式预测未知样本 testcase 的情感类别:

$$\text{predict}(\text{classifier}_1, \text{testcase}) = \begin{cases} \text{positive} & \text{positive} = \text{positive} + 1 \\ \text{neutral} & \text{neutral} = \text{neutral} + 1 \end{cases} \quad (21)$$

$$\text{predict}(\text{classifier}_2, \text{testcase}) = \begin{cases} \text{positive} & \text{positive} = \text{positive} + 1 \\ \text{negative} & \text{negative} = \text{negative} + 1 \end{cases} \quad (22)$$

$$\text{predict}(\text{classifier}_3, \text{testcase}) = \begin{cases} \text{neutral} & \text{neutral} = \text{neutral} + 1 \\ \text{negative} & \text{negative} = \text{negative} + 1 \end{cases} \quad (23)$$

最终未知样本 testcase 的情感类别 label 为

$$\text{label} = \max(\text{positive}, \text{negative}, \text{neutral}) \quad (24)$$

3 实验和结果分析

3.1 实验数据集

如表1所示,实验数据来自于京东和淘宝商品的用户评论文本。对数据集进行去重以及预处理操作,最终获得21 091条评论文本,其中8 033条为情感极性为积极类别的样本,4 355条为情感极性为中性类别的样本,8 703条为情感极性为消极类别的样本。由于三个类别的样本数目存在差异,数据的不平衡性会导致实验过程中分类器训练的结果严重偏向训练样本多的类别,对模型性能造成较大损失。因此,本文选取训练集时保证了三种类别的训练样本数量一致。

表1 语料集合样例

情感类别	样例文本
积极	a) 性价比很高,质量也很不错,真的是赚到了
	b) 高分屏,运气不错!性能还不知道,过几天装个游戏。外观还可以
	c) #蒙牛纯牛奶# 新包装感觉还不错
中性	a) 题目好,但内容空,给人感觉在说空话
	b) 我觉得本本还是不错的,但没想到买到康柏的,虽然现在是HP的,但还是有一点失望
	c) 由于是寄回老家的没有看到东西,但听家里人说还不错,还没有安装。有待后期追加评论
消极	a) 总的感觉前言不搭后语,浪费了银子,呜呜
	b) 实在一般,这样的酒店还要这个价钱,令人费解
	c) 太太太恶心了...蒙牛,去死

表2所示为训练集和测试集的划分。实验中对数据集采取欠采样方法生成训练集,即随机抽取积极、消极、中性三个类别的样本各4 000条,总计12 000条评论文本组成训练集,进而将数据集中的剩余评论文本作为测试集。

表2 实验数据集语料分布

样本类型	训练集	测试集	样本类型	训练集	测试集
积极	4 000	4 033	中性	4 000	355
消极	4 000	4 703	总数量	12 000	9 091

3.2 评价指标

本文选取准确率 P 、召回率 R 以及 F 值作为评价指标。如表3所示, P 表示对测试集进行情感分类后,预测为某个类别的样本中真正类别的样本所占的比例,即

$$P = \frac{r}{r+t} \quad (25)$$

如表3所示, R 表示对测试集进行情感分类后,预测为某个类别中的真实类别占所有真实类别的比例,即

$$R = \frac{r}{r+s} \quad (26)$$

表3 分类判别混淆矩阵

真实结果	预测结果	
	属于类别 A	不属于类别 A
属于类别 A	r	s
不属于类别 A	t	z

为了对准确率 P 和召回率 R 进行综合考虑,本文使用两者的加权调和平均数 F 来衡量最终分类效果。

$$F = \frac{2 \times P \times R}{P + R} \quad (27)$$

然而本文的情感分类任务面向的是三分类,因此在计算出各个类别所对应的准确率 P 和召回率 R 后,采用三个类别所对应的平均准确率 AVP 、平均召回率 AVR 以及平均 F 值 AVF 作为衡量情感分类器性能的评价指标。

3.3 实验模型参数设置

WEFF-BILSTM 神经网络模型训练的参数和函数设置如表4所示。网络模型采用 \tanh 作为激活函数,采用 Adam 作为梯度更新方法,且由于数据量偏少,所以将神经网络迭代次数设置为30。同时为了防止过拟合,利用 dropout 对网络参数进行约束。对于词向量维度、词频阈值以及窗口大小三个参数则采用单因子变量法实验确定最优值,实验结果如图4~6所示。

表4 实验参数设置

参数	参数/函数名称	参数值/函数值
-	激活函数	\tanh
-	优化方法	Adam
-	dropout	0.5
I	神经网络迭代次数	30
V	词向量维度	需对比实验[50, 200]
W	窗口大小	需对比实验[5, 50]
D	词频阈值	需对比实验[10, 100]

如图4所示,随着词向量维度从最初的50渐渐增大, AVF 值也在不断提升,直到词向量维度达到150时, AVF 值达到最优,进而随着词向量维度继续增加, AVF 值反而逐渐减小。因此,在神经网络训练时,选择150作为参数词向量维度 V 的值。

如图5所示,参数窗口大小 W 初值为5,随着 W 不断增大, AVF 值也在缓慢提升,直到 W 达到20时, AVF 值达到最优;随后尽管 W 不断增大, AVF 值始终未能超过前面的最优值。故在神经网络训练时,选择20作为参数窗口大小 W 的值。

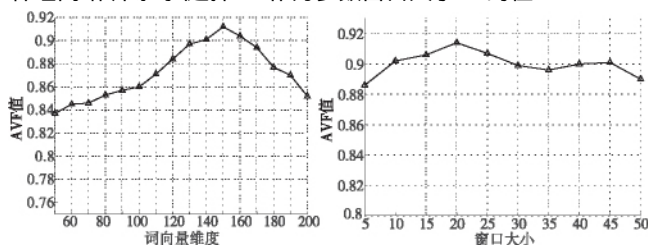


图4 基于词向量维度的 AVF 值趋势变化

图5 基于窗口大小的 AVF 值趋势变化

如图6所示,随着词频 D 值不断增大, AVP 值也在逐步提升,直到 D 值为60时达到最优值,随后 AVP 值随着 D 值的增大而减小。因此,在神经网络训练时,选择60作为参数词频阈值 D 的值。

3.4 对比实验

为验证本文提出的基于 word2vec 和双向 LSTM 的情感分类方法的有效性,将本文方法与 N-gram 词袋模型、GloVe 模型进行对比实验。具体方法定义如下:

a) 方法1 基于 uni-gram 和 bi-gram 模型构造特征向量,使用 SVM 对其进行情感分类。

b) 方法2 基于 word2vec 模型构建词向量,使用 SVM 进行情感分类。

c) 方法3 基于 uni-gram 和 bi-gram 模型构造特征向量,使用 logistic regression 对其进行情感分类。

d) 方法4 基于 word2vec 模型构造词向量,使用 logistic regression 对其进行情感分类。

e) 方法5 基于 uni-gram 和 bi-gram 模型构建特征向量,使用朴素贝叶斯进行情感分类。

f) 方法6 基于 word2vec 模型构建词向量,使用朴素贝叶斯进行情感分类。

g) GloVe-LSTM 基于 GloVe 模型训练词向量,网络模型结

构与本文提出的 WEEF-BiLSTM 模型结构相同,不过双向 LSTM 层改为单向 LSTM,分类层采用 softmax。

h) GloVe-BiLSTM 基于 GloVe 模型训练词向量,网络模型结构与本文提出的 WEEF-BiLSTM 模型结构相同,分类层采用 softmax。

i) word2vec-BiLSTM 基于 word2vec 模型训练词向量,网络模型结构与本文提出的 WEEF-BiLSTM 模型结构相同,分类层采用 softmax。

j) 本文方法 采用本文提出的深度词向量特征,使用 one-versus-one SVM 进行情感分类。

3.4.1 模型训练

基于以上介绍的 10 种方法来针对训练集进行模型的训练,具体的对比数据如下所示。使用方法 1~6 针对训练集进行实验所得到的对比结果如图 7 所示。可以明显发现,通过将词向量表示与传统机器学习算法结合可以有效地促进情感分类器性能的提升,证明了词向量表示方法能够在更抽象的层面上描述原始输入数据的特征分布情况。

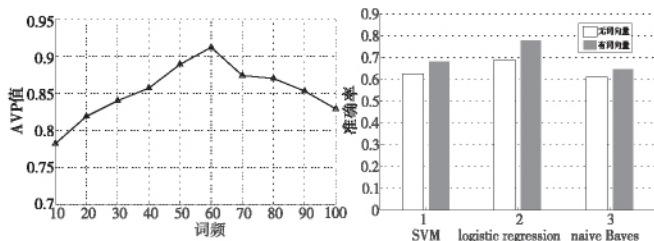


图 6 基于词频的 AVP 值趋势变化

图 7 基于 word2vec 的传统机器学习算法性能对比

针对训练集,使用本文方法与方法 2、4、6 进行对比实验,结果如图 8 所示。可以明显发现,使用基于 LSTM 的神经网络提取深度特征向量,降低了输出特征向量的维数,避免了维数灾难的出现,并且也比较明显地提升了情感分类器的性能,同时证明了 LSTM 神经网络结构对于网络文本进行情感分类具有健壮性。

针对训练集,使用本文方法与 GloVe-LSTM、GloVe-BiLSTM、word2vec-BiLSTM 三种方法进行对比实验,实验结果如图 9 所示。通过对比 GloVe-LSTM 和 GloVe-BiLSTM 发现,采取双向 LSTM 神经网络更能在深度层面上捕捉文本的上下文关系以及其中蕴涵的情感信息,进而促进情感分类器性能的提升;通过对比 GloVe-BiLSTM 和 word2vec-BiLSTM 发现,采取 GloVe 和 word2vec 两种模型的词向量表示方法,基于双向 LSTM 神经网络进行训练,所取得的情感分类器性能极其接近;通过对比 word2vec-BiLSTM 和本文方法发现,两者在训练集样本数为 4 000 以内时性能比较接近,但随着训练样本的增多,本文方法所取得的准确度明显比 word2vec-BiLSTM 高,证明了使用 one-versus-one SVM 分类器取代 softmax 分类器进行模型训练和情感分类时能够提高情感分类器的准确度。

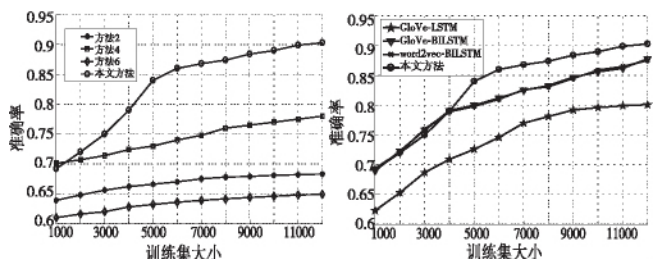


图 8 本文方法与传统机器学习方法性能对比

图 9 基于词向量与 LSTM 的网络模型性能对比

3.4.2 模型评估

基于上述 10 种方法得到的模型针对测试集进行实验进而

针对评价指标 AVP、AVR、AVF 进行对比分析,具体数据如表 5 所示。可以发现,本文方法所取得的准确率最高,达到 0.903,且 AVF 值也达到最高,为 0.868。以上对比实验结果证明,本文提出的基于 word2vec 和双向 LSTM 的情感分类模型能够有效地促进情感分类器性能的提升。

表 5 不同方法的分类性能对比

方法	AVP	AVR	AVF	方法	AVP	AVR	AVF
方法 1	0.625	0.590	0.601	方法 6	0.649	0.654	0.651
方法 2	0.683	0.664	0.673	GloVe-LSTM	0.797	0.742	0.754
方法 3	0.690	0.671	0.665	GloVe-BiLSTM	0.878	0.839	0.858
方法 4	0.782	0.753	0.767	word2vec-BiLSTM	0.875	0.846	0.860
方法 5	0.612	0.623	0.617	本文方法	0.903	0.837	0.868

4 结束语

情感分类作为自然语言处理领域的重要研究方向之一,应用传统机器学习算法以及目前流行的神经网络算法在人工特征工程建立以及情感语义理解等方面存在信息缺失,故本文针对网络文本基于 word2vec 和双向 LSTM 神经网络提出了一种用于情感分类的深度模型 WEEF-BiLSTM,将传统的机器学习算法与神经网络结合了起来。首先基于 word2vec 进行词嵌入从而获取词向量表示,进而通过本文设计的 WEEF-BiLSTM 神经网络模型提取文本的深度词向量特征,最后使用 one-versus-one SVM 进行情感分类。实验结果表明本文方法能够有效地对网络文本进行情感分类,且在一定程度上促进了情感分类器性能的提升。下一步将针对具有复杂结构以及特殊标签的文本进行情感分类的深入研究。

参考文献:

- [1] 陈龙,管子玉,何金红,等.情感分类研究进展[J].计算机研究与发展,2017,54(6):1150-1170. (Chen Long, Guan Ziyu, He Jinhong, et al. A survey on sentiment classification[J]. Journal of Computer Research and Development, 2017, 54(6): 1150-1170.)
- [2] 张向阳,那日萨.基于复杂网络的情感分类特征选择[J].计算机应用研究,2017,34(4):1000-1003. (Zhang Xiangyang, Na Risa. Emotional classification feature selection based on complex network[J]. Application Research of Computers, 2017, 34(4): 1000-1003.)
- [3] Li Shoushan, Huang Lei, Wang Rong, et al. Sentence-level emotion classification with label and context dependence[C]//Proc of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2015: 1045-1053.
- [4] 陈翠平.基于深度信念网络的文本分类算法[J].计算机系统应用,2015,24(2):121-126. (Chen Cuiping. Text classification algorithm based on deep belief network[J]. Computer Systems & Applications, 2015, 24(2): 121-126.)
- [5] Kim Y. Convolutional neural networks for sentence classification[C]//Proc of Conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.
- [6] Li Dong, Wei Furu, Tan Chuanqi, et al. Adaptive recursive neural network for target-dependent Twitter sentiment classification[C]//Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 49-54.
- [7] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[C]//Proc of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 655-665.
- [8] Dos Santos C, Gatti M. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proc of the 25th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2014: 69-78.

(下转第 3596 页)

5 结束语

本文针对新闻内容和新闻分类会对新闻推荐系统的推荐性能产生影响的情况下,提出了一种基于 VSM 和 bisecting K-means 聚类的用户兴趣模型构建方法;考虑到用户兴趣会随时间而发生漂移,在该模型基础上,参考艾宾浩斯遗忘曲线构造了遗忘函数,并利用遗忘函数对该模型进行了时间加权更新。实验结果表明,本文构建的模型与协同过滤算法相比,推荐性能有所提高;更新后的模型与原始模型相比,推荐性能又有进一步改善,但提升较少。在实验数据中, F 值一直呈现先高后低的趋势,这一现象目前还难以解释,需进一步深入研究。总的来看,本文提出的方法可作为新闻推荐领域的一种方案,在模型构建与利用遗忘函数更新模型方面可作为研究参考。下一步将研究用户兴趣模型更新的问题,以及对结合地理位置信息的新闻推荐方法进行研究。

参考文献:

- [1] 冷亚军,陆青,梁昌勇.协同过滤推荐技术综述[J].模式识别与人工智能,2014,27(8):720-734. (Leng Yajun, Lu Qing, Liang Changyong. Survey of recommendation based on collaborative filtering [J]. Pattern Recognition & Artificial Intelligence, 2014, 27(8): 720-734.)
- [2] 孔艳莉.基于协同过滤算法的个性化推荐技术研究[D].北京:北京工业大学,2016. (Kong Yanli. Research on personalized recommendation based on collaborative filtering [D]. Beijing: Beijing University of Technology 2016.)
- [3] 吴彦文,齐旻,杨锐.一种基于改进型协同过滤算法的新闻推荐系统[J].计算机工程与科学,2017,39(6):1179-1185. (Wu Yanwen, Qi Min, Yang Rui. New recommendation system based on an improved collaborative filtering algorithm [J]. Computer Engineering & Science 2017, 39(6): 1179-1185.)
- [4] Garcin F, Zhou Kai, Faltings B *et al.* Personalized news recommendation based on collaborative filtering [C]//Proc of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. Washington DC: IEEE Computer Society 2012: 437-441.
- [5] 彭菲菲,钱旭.基于用户关注度的个性化新闻推荐系统[J].计算机应用研究,2012,29(3):1005-1007. (Peng Feifei, Qian Xu. Personalized recommendation system for news based on user concern [J]. Application Research of Computers 2012, 29(3): 1005-1007.)
- [6] Okura S, Tagami Y, Ono S *et al.* Embedding-based news recommendation for millions of users [C]//Proc of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press 2017: 1933-1942.
- [7] Zhang Kuai, Xin Xin, Luo Pei *et al.* Fine-grained news recommendation by fusing matrix factorization, topic analysis and knowledge graph representation [C]//Proc of IEEE International Conference on Systems, Man and Cybernetics. Piscataway, NJ: IEEE Press, 2017: 918-923.
- [8] 古万荣,董守斌,何锦潮,等.基于二次聚类的新闻推荐方法[J].华南理工大学学报:自然科学版,2014,42(7):15-20,32. (Gu Wanrong, Dong Shoubin, He Jinchao *et al.* A news recommendation method based on two-fold clustering [J]. Journal of South China University of Technology: Natural Science Edition 2014, 42(7): 15-20, 32.)
- [9] 李佳珊.个性化新闻推荐引擎中新闻分组聚类技术的研究与实现[D].北京:北京邮电大学,2013. (Li Jiashan. Research and implementation of text clustering for personalized news recommendation system [D]. Beijing: Beijing University of Posts & Telecommunications, 2013.)
- [10] 费洪晓,戴弋,穆琨,等.基于优化时间窗的用户兴趣漂移方法[J].计算机工程,2008,34(16):210-211,214. (Fei Hongxiao, Dai Yi, Mu Jun *et al.* Method of drifting user's interests based on time window optimization [J]. Computer Engineering, 2008, 34(16): 210-211, 214.)
- [11] 成伟丹.基于遗忘函数和用户的协同过滤推荐算法研究[D].杭州:浙江工业大学,2016. (Cheng Weidan. Research on collaborative filtering algorithm based on forgetting function and user [D]. Hangzhou: Zhejiang University of Technology 2016.)
- [12] Sun Baoshan, Dong Lingyu. Dynamic model adaptive to user interest drift based on cluster and nearest neighbors [J]. IEEE Access, 2017, 5: 1682-1691.
- [13] 邢春晓,高凤荣,战思南,等.适应用户兴趣变化的协同过滤推荐算法[J].计算机研究与发展,2007,44(2):296-301. (Xing Chunxiao, Gao Fengrong, Zhan Sinan *et al.* A collaborative filtering recommendation algorithm incorporated with user interest change [J]. Journal of Computer Research and Development 2007, 44(2): 296-301.)
- [14] 姚清耘.基于向量空间模型的中文文本聚类方法的研究[D].上海:上海交通大学,2008. (Yao Qingyun. Research of VSM-based Chinese text clustering algorithms [D]. Shanghai: Shanghai Jiao Tong University 2008.)
- [15] Abuaiaad D. Using bisect K-means clustering technique in the analysis of Arabic documents [J]. ACM Trans on Asian and Low-Resource Language Information Processing, 2016, 15(3): article No. 17.
- [16] Zhuang Yu. Symmetric repositioning of bisecting K-means centers for increased reduction of distance calculations for big data clustering [C]//Proc of IEEE International Conference on Big Data. Piscataway, NJ: IEEE Press 2016: 2709-2715.
- [17] Ebbinghaus H. Memory: a contribution to experimental psychology [J]. Annals of Neurosciences 2013, 20(4): 155-156.
- [18] 项亮.推荐系统实践[M].北京:人民邮电出版社,2012: 44-59. (Xiang Liang. Recommended system practice [M]. Beijing: Posts & Telecom Press 2012: 44-59.)

(上接第3587页)

- [9] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]//Proc of Conference on Neural Information Processing Systems. 2012: 1097-1105.
- [10] Zhu Xiaodan, Sobhani P, Guo Hongyu. Long short-term memory over recursive structures [C]//Proc of the 32nd International Conference on International Conference on Machine Learning. New York: ACM Press 2015: 1604-1612.
- [11] Ren Yafeng, Wang Ruimin, Ji Donghong. A topic-enhanced word embedding for Twitter sentiment classification [J]. Information Science 2016, 369(11): 188-198.
- [12] Chen Tao, Xu Ruifeng, He Yulan. Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN [J]. Expert Systems with Applications 2017, 72(4): 221-231.
- [13] 任勉,甘刚.基于双向 LSTM 模型的文本情感分类[J].计算机工程与设计,2018,39(7):2064-2068. (Ren Mian, Gan Gang. Text sentiment classification based on two-way LSTM model [J]. Computer Engineering and Design 2018, 39(7): 2064-2068.)
- [14] Mikolov T, Sutskever I, Chen Kai *et al.* Distributed representations of words and phrases and their compositionality [C]//Proc of the 26th International Conference on Neural Information Processing Systems. [S. l.]: Curran Associates Inc. 2013: 3111-3119.
- [15] Greff K, Srivastava R K, Koutník J *et al.* LSTM: a search space odyssey [J]. IEEE Trans on Neural Networks & Learning Systems, 2017, 28(10): 2222-2232.
- [16] 李婷婷,姬东鸿.基于 SVM 和 CRF 多特征组合的微博情感分析[J].计算机应用研究,2015,32(4):978-981. (Li Tingting, Ji Donghong. Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features [J]. Application Research of Computers 2015, 32(4): 978-981.)
- [17] 王汝娇,姬东鸿.基于卷积神经网络与多特征融合的 Twitter 情感分类方法[J].计算机工程,2018,44(2):210-219. (Wang Ru-jiao, Ji Donghong. Twitter sentiment classification method based on convolutional neural network and multi-feature fusion [J]. Computer Engineering 2018, 44(2): 210-219.)
- [18] 殷昊,李寿山,贡正仙,等.基于多通道 LSTM 的不平衡情绪分类方法[J].中文信息学报,2018,32(1):139-145. (Yin Hao, Li Shoushan, Gong Zhengxian *et al.* Imbalanced emotion classification based on multi-channel LSTM [J]. Journal of Chinese Information Processing 2018, 32(1): 139-145.)