

基于深度学习的情感分类技术在高校舆情分析中的应用研究

黄萍, 朱惠娟, 陈琳琳

(南京理工大学紫金学院, 江苏 南京 210000)

✉huangping984@njjust.edu.cn; elainezhj@qq.com; chenlinlin606@njjust.edu.cn



摘要: 传统机器学习的自然语言处理系统特别依赖人工手动标记的特征, 极其耗时且容易出现维度爆炸等难以解决的问题。本文采用基于卷积神经网络(CNN)的深度学习技术来解决这一问题。通过收集校园热点话题进行预处理以及运用Word2vec模型生成词向量后, 运用卷积神经网络提取其中的特征并进行情感倾向分类。通过实验数据的比较, 基于卷积神经网络(CNN)的情感倾向分类获得了89.76%的准确率, 较传统的支持向量机(SVM)提高了7.3%, 获得更好的分类性能。本文的研究对高校治理能力和治理体系现代化建设具有积极作用。

关键词: 自然语言处理; 卷积神经网络; 情感倾向分析; 舆情分析

中图分类号: TP39 **文献标识码:** A

Application of Emotion Classification Technology based on Deep Learning in University Public Opinion Analysis

HUANG Ping, ZHU Huijuan, CHEN Linlin

(Zijin College, Nanjing University of Science and Technology, Nanjing 210000, China)

✉huangping984@njjust.edu.cn; elainezhj@qq.com; chenlinlin606@njjust.edu.cn

Abstract: Traditional natural language processing systems for machine learning rely heavily on manually marked features, which are extremely time-consuming and prone to difficult problems like dimensional explosions. This paper proposes to use CNN-based (Convolutional Neural Network) deep learning technology to solve this problem. After hot topics on campus are collected for preprocessing and generating word vectors using word2vec model, CNN is used to extract features and classify emotional tendencies. Through experimental comparison, the emotion tendency classification based on CNN has an accuracy of 89.76%, which is 7.3% higher than that of traditional Support Vector Machine (SVM) and has better classification performance. This research plays a positive role in the modernization of university governance ability and governance system.

Keywords: natural language processing; convolutional neural network; emotion tendency analysis; public opinion analysis

1 引言(Introduction)

随着信息技术的迅速发展和自媒体的普及, 网络对大学生的思维方式、思想观念、人际交往和学习生活产生了深刻影响, 各个高校校园文化的展示不再局限于校园内部, 各种虚拟网络平台也成为校园文化和展示的平台。借助自媒体平台, 学生们可以随时随地在社交网络上发表自己的观点和见解, 而且这些观点和见解往往是带有明显的情感倾向

的, 在一定程度上, 这些正面或负面的高校网络舆情也客观地反映出校园文化的健康程度。如何在海量的数据中捕获到用户的情感倾向信息, 挖掘出带有情绪和喜恶的主观信息, 是情感倾向分类要做的主要工作。情感倾向分类可以对文本所表达的带有主观情感色彩的信息进行处理、挖掘, 并分析其中包含的积极或消极信息, 通过判断信息的情绪极性进行舆情态势感知和预警, 有助于对极端情绪的检测与控制。

总之,在现代高校管理中,充分挖掘师生对热点舆情事件的情感倾向,分析其所表达价值取向或者事件产生的深层次原因,对开展校园网络舆情研究和进行有针对性的学生思想引导工作是至关重要的,对推动网络空间的科学治理也起到促进作用。

2 基于深度学习的高校网络舆情分析系统 (University network public opinion analysis system based on deep learning)

情感分类算法研究是网络舆情分析的一个重要研究领域,对于舆情分析有着重要的意义。近年来,国内高校网络舆情突发事件频繁发生,比如2020年的“山西作弊大学生坠亡”“疫情期间高校施行‘相对封闭式管理’”等。这些事件所爆发出的网络舆论给相关高校造成了极大的困扰。因此,在网络空间科学治理工程的背景下,分析和研究高校网络舆情发展和传播规律,探索如何在高校师生中开展有效的网络舆情管理和引导已成为需要深入研究思考的问题。作为高校,面对当下日益复杂以及多元化的网络环境,要做好网络舆情的预警工作,运用计算机辅助技术实时收集网络舆情数据,对其中的热点话题数据进行分析研判,精确地发现引发舆情危机的节点,在短时间内制定有针对性的处置策略,不给舆情危机发酵的时间和空间^[1]。因此,若能对网络热点话题或事件进行搜索和分析,并总结出其中正面信息和负面信息的比例,进而对一些学生关注度高的问题及时进行解决以及疏导,这对于完善高校治理无疑是非常有用的。

在国内,基于深度学习的文本情感分类研究起步较晚,但发展迅猛,目前已经有很多研究成果涌现出来。刘龙飞等人^[2]使用CNN方法对微博文本的情感进行研究,其中原始特征由字向量与词向量同时构成,在COAE2014上取得不错的效果。刘智鹏等人^[3]构造与设计了CNN与RNN模型,并进行了有效的融合,利用各自对短文本的处理优势进行商品的评价分类,获得了较好的文本情感识别性能。周锦峰等人^[4]通过堆叠多个卷积层,提取不同窗口的局部语义特征以及基于全局最大池化层构建分类模块,获得了较快的文本情感分类速度。蔡庆平等人^[5]设计了基于Word2vec和CNN的产品评论细粒度情感分析模型,有效地发现用户对产品特征的关注度和满意度。

本文运用基于深度学习的情感分析技术手段,分析和研判网络中高校热点话题评论中所蕴含的情感倾向信息,并进行网络舆情监测。网络舆情分析分为舆情信息采集、文本数据预处理、词向量化、舆情数据学习及分析、舆情预警(结果可视化)五个步骤。首先利用网络爬虫技术完成数据的收集;接着对数据进行中文分词、去停用词操作,保留语句中的关键信息;再运用词向量工具将词转换成词向量,以便可以被卷积神经网络学习,通过网络的学习,提取其中的特征,最终可被用于情感极向的分类,如图1所示。可视化模块则用于

显示分类结果,负面评论达到一定比例时,需要对相关问题进行疏导。

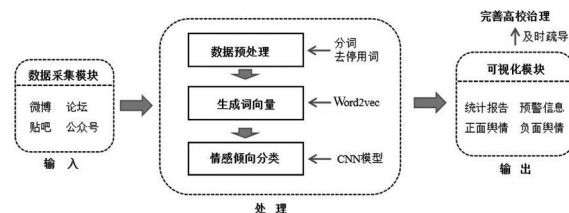


图1 高校舆情设计系统框架图

Fig.1 Framework of public opinion design system in colleges and universities

2.1 数据采集模块

为了能够快速准确地获取最新的网络舆情数据,本文利用分布式网络爬虫对指定网站进行数据爬取,简单清洗之后,作为系统实验数据来源。首先将数据收集任务分解成多个子任务,分配给多个爬虫线程来共同完成;接着通过向网站的服务器发送请求,获取网页源代码并进行数据清洗、去重去噪,将一些标签、CSS代码内容、空格字符、脚本标签等内容处理掉,使冗余的网页数据变得结构清晰^[6];最终将这些信息存储为纯文本数据,为接下来的数据分析和处理提供基础。

2.2 数据预处理

通过网络爬虫获取的纯文本数据需要转化为适合于表示和分类的干净的词序列。由于中文句子中的词语之间没有明确的分隔符且存在一定的噪音信息,因此在预处理阶段要对句子进行分词、去除停用词等操作。

(1)分词。中文分词是文本处理的一个基础步骤,由于中文句子不像英文句子那样词与词之间有明显的分隔符,因此需要利用中文分词技术将词语切分开。成熟的中文分词算法能够达到更好的自然语言处理效果,帮助计算机理解复杂的中文句子。本文采用基于词典分词的jieba分词器,它运用有向无环图的查找算法,通过动态规划,从后至前使得词的切割组合联合概率最大。对于不在词典里的词再使用HMM算法来进行二次分词,采用分词中的序列标注方法,使用模型识别词每个位置的状态值^[7]。

(2)去停用词。通过分词可以把句子分出很多词语,但是其中有些词未包含实际含义,如“的”“了”“着”等,还有一些英文字符、数字、标点符号等。这些词普遍存在,又未包含具体含义,同时记录它们需要较大的空间。本文根据网上现有资源,对“哈工大停用词词库”“百度停用词表”等多种停用词表合并整理后,生成了一个共有1,598个停用词的停用词表。在分词过程中,判断得到的每个中文词是否是停用词,如果是停用词则直接删除,以便降低特征的维度,提高关键词密度。

2.3 文本的分布式表示

预处理后的文本是一种计算机无法直接处理的非结构化数据,需要转换成结构化数据——向量。本文采用Word2vec

词向量工具将文本转换成词向量, 以便于网络学习。Word2vec是MIKOLOV等人^[8]提出的一种文本分布式表示方法, 由此词嵌入的思想开始应用到自然处理的领域。它是一款将词表征为实数值向量的高效工具, 背后的模型是CBOW或者Skip-gram, 使用了Hierarchical Softmax或者Negative Sampling的优化方法^[9]。Word2vec能够将每个词映射成一个K维的实数向量, 精确地度量词与词之间的关系, 挖掘词与词之间的联系。

本次实验采用CBOW模型进行词向量表示, 通过输入特征词的上下文相关词对应的词向量来预测输出特征词的词向量。用CBOW模型训练词向量, 首先需要根据语料建立一张词汇表, 并给表中的每个词语生成随机的词向量; 然后将特定词的上下文词向量输入CBOW, 再由隐含层进行累加, 到第三层中的哈夫曼树, 沿着特定的路径到达叶子节点, 从而完成对特定词语的预测, 训练结束后就可以从词汇表中得到每一个词语所对应的词向量。

2.4 深度学习情感分类模型

本文采用卷积神经网络模型来解决中文情感倾向分析问题, 将由Word2vec转化后的词向量矩阵作为卷积神经网络的输入; 然后通过卷积层进行特征提取, 再用最大池化法降低每条评论特征向量的维度; 最后在全连接层由ReLU函数做出分类输出, 将评论信息分成积极和消极两种。卷积神经网络是一种多层的监督学习神经网络, 由输入层(Input Layer)、卷积层(Convolution Layer)、池化层(Pooling Layer)、全连接层(Fully Connected Layer)和输出层(Output Layer)组成, 其中卷积层和池化层是实现特征提取功能的核心模块, 结构如图2所示。

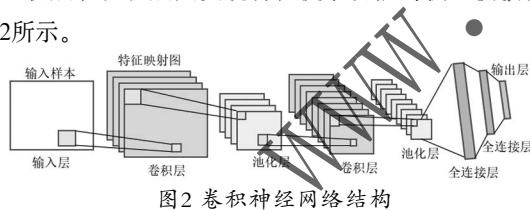


图2 卷积神经网络结构

Fig.2 Structure of Convolution Neural Network

卷积层: 在第一层卷积层中对输入的词向量矩阵进行卷积运算后, 可以得到对应的特征图。卷积运算使原信号特征增强的同时, 还可以降低噪音, 提取输入样本中的不同特征。同时, 一个卷积层中可以有多个不同的卷积核, 每一个卷积核都对应一个特征图, 如图3所示。

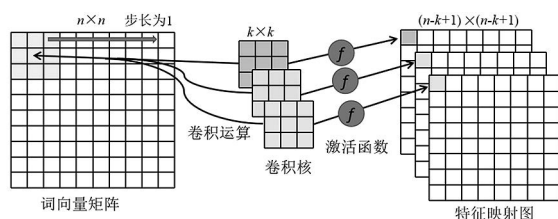


图3 具有多个卷积核的卷积层

Fig.3 Convolution layer with multiple convolution kernels

池化层: 经过卷积运算后的特征矩阵尺寸往往比较大, 运用池化(Pooling)操作可以减小卷积层产生的词向量矩阵的维度, 使得参数的数量和计算量下降。本实验采用最大池化法来降维, 将卷积运算后的特征矩阵划分为若干个矩形区域, 输出每个子区域最大值, 减小数据的空间大小, 如图4所示。通过降低特征矩阵的维度, 使得特征表示对输入词向量的位置变化具有更好的稳健性, 还在一定程度上预防过拟合。



图4 最大池化

Fig.4 Maximum pooling

全连接层: 它是整个卷积神经网络中的“分类器”。全连接层起到将学到的“分布式特征表示”映射到样本标记空间的作用, 灰色的神经元表示这个特征被找到了(激活了), 得到的激活值即卷积神经网络提取到的特征, 如图5所示。

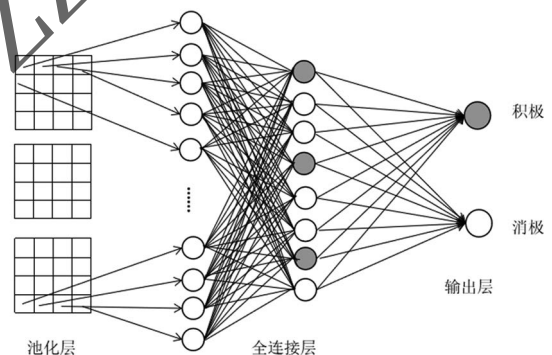


图5 全连接层的输入与输出

Fig.5 Input and output of full connection layer

3 实验与结果分析(Experiment and result analysis)

3.1 实验数据来源

高校网络舆情是建立在大众网络舆情基础之上的, 其主要内容基本都是反映高校学习生活中的师生关系或者学习生活的一些典型事件。本文实验所用到的数据均为使用爬虫软件, 以“封闭式校园管理”“考研扩招”“直播授课”“旷课”“退学”“学术不端”“就业”等关键字在微博网站中爬取到的2020年1—12月的15,000条相关评论文本数据, 其中90%的数据作为训练集(train), 10%的数据作为评估集(val)。

3.2 实验结果分析

本文基于Keras构建了用于进行文本的情感倾向分析的CNN网络。网络参数设置如表1所示。

表1 CNN模型训练参数设置

Tab.1 CNN model training parameter setting

参数名	参数值
词嵌入维度	100 × 100
CNN卷积核个数	128
CNN卷积核大小	5
全连接层个数	100
输出层个数	1
模型迭代次数	10
每批训练集的数据量	128

使用CNN模型处理数据过程中,迭代10次之后基本可以达到较好的效果,其收敛情况如图6所示。通过模型在train/val集上的准确率(acc)、损失函数(loss)的计算,从而更新模型参数,减小优化误差(Optimization Error),即在损失函数与优化算法的共同作用下,减小模型的经验风险,同时对模型的效果进行度量。一般来说loss越小,表示网络优化程度越高,acc就会越高。

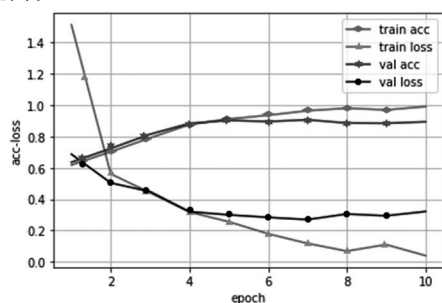


图6 CNN网络的acc曲线和loss曲线

Fig.6 Acc curve and loss curve of CNN network

本文运用Word2vec+SVM和Word2vec+CNN这两种分类方法对爬取到的高校相关舆情信息进行积极和消极类别的情感倾向分类。两种算法均采用100维词向量,对它们的准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1值(F1 Score)进行比较分析,找出比较优异的分类方法,如表2所示。

表2 实验结果比较

Tab.2 Comparison of experimental results

模型	准确率	精确率	召回率	F1值
Word2vec+CNN	89.76%	91.80%	91.48%	91.66%
Word2vec+SVM	82.46%	95.44%	75.23%	84.14%

通过使用同样的舆情数据进行实验,Word2vec+CNN模型与Word2vec+SVM模型相比,SVM在挖掘词向量中隐藏的分类特征过程中,会损失词与词之间的语义信息;而CNN却可以提取不同维度的特征,进而更好地挖掘出舆情语料词向量中隐藏的特征信息。因此,Word2vec+CNN模型在准确率、召回率和F1值上都比Word2vec+SVM方法具有更加良好的分类效果。

4 结论(Conclusion)

本文描述了基于Word2vec+CNN的情感倾向分类模型在校园舆情分析与监测方面的应用研究。在收集了2020年校园热点舆情数据,并进行良好的数据预处理的基础上,运用Word2vec工具将网络舆情文本数据转化为词向量,再分别使用CNN模型和SVM模型对词向量进行情感倾向分类。总体来说,在本次分类任务中CNN模型获得了比SVM模型更好的性能表现。

另外,在分类过程中还发现,文本特征的数量、数据标注精确度对实验的精度和时间也有一定的影响。然而,网络舆情数据由于数据量大、杂乱无章且无标签,存在依赖人工标注的成本高,效率低的问题。因此在未来工作中,应在特征选择上增加研究投入,同时结合无监督数据的特征学习和半监督学习自动标注方法进行情感分类研究,以期能够显著提升分类研究的最终结果精度,精确挖掘网络中的学生情绪动向,完善高校治理,对治理能力和治理体系现代化建设起到积极作用。

参考文献(References)

- [1] 孔晓虹.新媒体环境下高校应对网络舆情的探索[J].中国高等教育,2020(Z1):67-68.
- [2] 刘龙飞,杨亮,张绍武,等.基于卷积神经网络的微博情感倾向性分析[J].中文信息学报,2015,29(6):159-165.
- [3] 刘智鹏,何中市,何伟东,等.基于深度学习的商品评价情感分析与研究[J].计算机与数字工程,2018,46(5):921-927.
- [4] 周锦峰,叶施仁,王晖.基于深度卷积神经网络模型的文本情感分类[J].计算机工程,2019,45(3):300-308.
- [5] 蔡庆平,马海群.基于Word2Vec和CNN的产品评论细粒度情感分析模型[J].图书情报工作,2020,64(6):49-58.
- [6] 朱琪.基于网络爬虫的舆情分析预警系统设计[J].电子设计工程,2020,28(22):56-60.
- [7] 祝永志,荆静.基于Python语言的中文分词技术的研究[J].通信技术,2019,52(7):1612-1619.
- [8] MIKOLOV T, SUTSKEVER I, KAI C, et al. Distributed representations of words and phrases and their compositionality[J]. Advances in Neural Information Processing Systems, 2013(26):3111-3119.
- [9] 梁翼园.基于改进的Word2vec的评论情感倾向性分析[D].长沙:湖南大学,2019.

作者简介:

黄萍(1982-),女,硕士,讲师.研究领域:人工智能,图像处理,自然语言分析,计算机教育.

朱惠娟(1985-),女,硕士,副教授.研究领域:图像处理,人工智能,虚拟现实.

陈琳琳(1981-),女,硕士,副教授.研究领域:人工智能,图像处理,计算机教育.