情报学报　2020 年 1 月　第 39 卷　第 1 期
Journal of the China Society for Scientific and Technical Information, Jan. 2020, 39(1): 81-89

DOI: 10.3772/j.issn.1000-0135.2020.01.008

基于OCC模型和LSTM模型的财经微博文本情感分类研究

吴　鹏[1,2]，李　婷[1,2]，仝　冲[1,2]，沈　思[1,2]

（1. 南京理工大学经济管理学院，南京　210094；2. 江苏省社会公共安全科技协同创新中心，南京　210094）

摘　要　为了解决财经微博文本中网民情感状态转移的时序数据分析问题，本文提出一个基于认知情感评价模型（Ortony，Clore & Collins，OCC）和长短期记忆模型（long short term memory，LSTM）的财经微博文本情感分类模型（OCC-LSTM）。基于OCC模型从网民认知角度建立情感规则，对财经微博文本进行情感标注，并作为LSTM模型进行深度学习的训练集；基于LSTM模型，使用深度学习中的TensorFlow框架和Keras模块建立相应的实验模型，进行海量微博数据情感分类，并结合13家上市公司3年的微博文本数据进行实证研究和模型验证对比。实证研究结果发现本文提出的模型取得了89.45%的准确率，高于采用传统的机器学习方式的支持向量机方法（support vector machine，SVM）和基于深度学习的半监督RAE方法（semi-supervised recursive auto encoder）。

关键词　长短期记忆模型；OCC模型；财经微博；情感分类

Sentiment Classification of Financial Microblog Text Based on the Model of OCC and LSTM

Wu Peng[1,2], Li Ting[1,2], Tong Chong[1,2] and Shen Si[1,2]

(1. *School of Economics and Management, Nanjing University of Science & Technology, Nanjing*　210094;
2. *Jiangsu Collaborative Innovation Center of Social Safety Science and Technology, Nanjing*　210094)

Abstract: To analyze the time series data of sentimental status transformation of online users in financial microblog text, this paper proposed a model of sentiment classification of financial microblog text based on the Long Short Term Memory model combined with the OCC model. The rules of sentiment were proposed from the view of online users' cognition based on the OCC model, and these rules can be taken as training sets for the emotion annotation of the financial microblog texts in the process of deep learning based on the model of LSTM. The sentiment classification task was fulfilled by the Keras module of the TensorFlow framework based on the LSTM model. An experiment was carried out to attest to the utility of the proposed model using financial microblog texts from thirteen listed companies in the last three years. The findings showed that the proposed model achieved 89.45% accuracy, and the accuracy of the proposed model is better than that of the SVM method and the semi supervised RAE method.

Key words: Long Short Term Memory model; OCC model; financial microblog; sentiment classification

收稿日期：2019-01-02；**修回日期**：2019-03-01

基金项目：国家自然科学基金项目“突发事件网民负面情感的模型检测研究”（71774084），“社会化影响下个体信息认知处理中的扭曲与偏见机制研究”（71471089）；国家社会科学基金项目“基于社会网络分析的网络舆情主题发现研究”（15BTQ063）。

作者简介：吴鹏，男，1976年生，博士，教授，主要研究领域为网络舆情，E-mail：wupeng@njust.edu.cn；李婷，女，1996年生，硕士研究生，主要研究领域为情感分析；仝冲，男，1996年生，硕士研究生，主要研究领域为用户行为；沈思，女，1983年生，博士，讲师，主要研究领域为信息检索。

1 引 言

财经领域网络舆情对上市公司和其他相关研究机构有着至关重要的作用，而财经微博情感分类又是财经网络舆情分析中一个关键的环节。情感分类是通过从文本信息提取基于某个事件、主题、对象等客体内容的主观评价情绪和态度，并利用一定的理论和技术进行归纳和总结的方法[1]。

与传统文本的情感分类不同，微博有其独特的情感特征，需要分析字里行间的内在含义，需要从不同维度对微博信息的特征进行分析，否则很难准确判断情感倾向。同时，微博具有篇章短小精悍、语言结构口语化、存在表情符号和创造性语言的特征，增加了语言处理和分析的难度[2]。

已有微博情感分类研究通常分为基于规则的情感分类和基于机器学习的分析分类方法。基于规则的情感分类方法一般通过构建相应的规则库或利用专家评估的方法对情感进行分类。基于机器学习的情感分类方法又可以分为无监督和有监督两种方式，目前常用的方法是有监督的机器学习，如支持向量机、朴素贝叶斯和神经网络等。但在对财经领域网络舆情进行情感分类时，通常将股票和股票分析机构的新闻信息看作同类网民，在实际应用中难以区分二者的不同情感类型[3-4]，缺乏网民情感状态转移的时序数据分析[5]，在对大数据时代海量的、高时效性的财经微博时序数据分析上具有局限性[6]。

序列化模型则将文本看作是有序的词语序列，这种模型结合文本的有序性以及词语间的关联性，可以学习到一些词袋模型无法学习到的语义信息[7]。在解决序列化问题时，循环神经网络（recurrent neural network，RNN）被证明是一种有效的方法，可以充分地利用上下文信息，其中长短期记忆（long short term memory，LSTM）模型[8]是一种有效的链式循环神经网络，因为其能有效利用序列数据中长距离依赖信息的能力，被认为特别适合文本序列数据的处理。LSTM模型不依赖于句子的标签和形式，通过分层的方式增强词与词之间、句与句之间的联系，通过主题分类判断情感倾向，再将每一层结果进行加权求和得到最终的情感倾向，从而解决了因时间迁移导致数据模糊而无法计算的问题，增强了分类的准确性，更适用于口语化的中文微博，提高整条微博的识别性，被广泛用于语言模型、机器翻译、语音识别等领域。但由于LSTM模型是一种链式结构，不能有效表征语言的结构层次信息，进行网民情感的分类[7]。

进行网民情感状态转移的研究时，首先应该判断网民所处的情感状态，在财经舆情中，网民的情感源于其对事件的评价，并受到自身特征和外部舆情环境等多方面因素的影响，评价过程具有主观性，取决于网民的特定目标、信念和规范等，具有典型的离散型特征。1988年，Ortony、Clore、Collins在《情感的认知结构》提出的OCC模型正是依据评价的认知过程而产生的一种认知情感评价模型[9]。

本文基于LSTM模型和认知情感评价（OCC）模型构建财经微博文本情感分类模型，基于OCC模型从网民认知角度建立情感规则，对财经微博文本进行情感分类标注作为训练集，基于LSTM模型，构建海量微博数据情感分类模型，以提高财经微博网民情感倾向预测的准确率。从而能够即时有效地为财经领域的微博情感分析研究提供有力的决策支持。

2 情感分类研究综述

2.1 长短期记忆模型

1997年，Hochreiter等[8]提出了长短期记忆（LSTM）模型，用于解决循环神经网络（RNN）训练时的梯度爆炸和梯度消失问题，使得RNN能真正有效地利用长距离的序列信息。研究者基于LSTM模型，不断改进RNN的效能。2014年，Sutskever等[10]提出了多层LSTM模型框架，能够使更高层次的LSTM模型捕捉到更长距离的信息；2015年，Li等[11]还提出了层次的LSTM模型，使用该模型分别处理词、句子和段落级别输入，并使用自动编码器（auto encoder）来检测文档特征抽取和重建能力。

在社会化媒体情感倾向性识别方面，Nguyen等[12]提出用于Twitter文本情感分析的深度双向LSTM神经网络模型（deep bi-directional long short-term memory neural networks），Wang等[13]提出预测Twitter文本的LSTM模型，Cheng等[14]使用LSTM模型进行新浪微博文本的情感分析和观点提取，陈鹏[15]提出联合LSTM模型与卷积神经网络（convolutional neural networks，CNN）模型判定评价对象情感的方法。

已有研究表明，LSTM模型结构能够有效地进

词袋模型(Bag of Words，简称BoW)，即将所有词语装进一个袋子里，不考虑其词法和语序的问题，即每个词语都是独立的，把每一个单词都进行统计，同时计算每个单词出现的次数。也就是说，词袋模型不考虑文本中词与词之间的上下文关系，仅仅只考虑所有词的权重，而权重与词在文本中出现的频率有关。

词袋模型的三部曲：分词（tokenizing），统计修订词特征值（counting）与标准化（normalizing）。

词袋模型首先会进行分词，在分词之后，通过统计每个词在文本中出现的次数，我们就可以得到该文本基于词的特征，如果将各个文本样本的这些词与对应的词频放在一起，就是我们常说的向量化。向量化完毕后一般也会使用TF-IDF进行特征的权重修正，再将特征进行标准化。再进行一些其他的特征工程后，就可以将数据带入机器学习算法进行分类聚类了。

当然，词袋模型有很大的局限性，因为它仅仅考虑了词频，没有考虑上下文的关系，因此会丢失一部分文本的语义。但是大多数时候，如果我们的目的是分类聚类，则词袋模型表现的很好。

行文本处理，因此本文将选取LSTM模型作为财经微博文本情感分析的深度学习模型。

2.2　OCC模型

OCC模型情感分类基于3个评价标准：事件的结果（consequences of events）、对象的行为（action of agents）和对象的描述（aspects of objects）。基于这3个评价标准，根据强度不同或诱发原因不同将情感划分为22类，通过情感评价标准的树形结构形象全面展示不同类型的情感。因此，OCC模型是现今最详细的情感分类模型，被认为是第一个以计算机实现为目的而发展起来的情感模型，是计算机的缺省情感模型[16]，被国内外学者广泛研究和引用。其主要应用领域包括以下3个方面：①基于语法和语义角度，将OCC模型逻辑形式化，形成正式的情感模型[16]；②基于OCC模型，研究情感如何影响认知和行为决策，设计相关情感模型应用于机器人、虚拟游戏人物等领域的智能体[17]；③基于OCC模型，结合群体行为决策模型等研究个体情绪状态和认知状态的调节[18]。

研究者已经开始应用OCC模型进行社会化媒体的情感分类研究。例如，Bartneck[19]对OCC模型在具体人物的情感分类的应用进行了定量分析，并对原本22种情感分类模型进行了进一步的整合。Roberts等[20]则对Twitter进行了情感分类的研究，对Twitter文本进行了标注，并人工划分为愤怒、厌恶、恐惧、喜爱、悲伤、惊讶，通过定量分析的方式，训练了情感分类模型，对Twitter文本进行情感分析。

2.3　OCC模型与LSTM模型关联研究

Yalcin等[21]结合OCC模型和LSTM模型和进行智能体（agent）交互时的情感状态转移研究，Deng等结合OCC模型和LSTM模型进行信息检索中的用户情感分类研究[22]、深度情感建模研究[23]。

这些研究为本文的研究提供了理论方法支撑。本文在上述研究基础上，基于OCC模型和LSTM模型对财经微博信息进行情感分类研究，试图解决目前财经领域微博舆情情感分类缺乏实用性的问题，捕获文本更深层次的语义语法信息。

3　模型设计

本文的研究是为了将财经微博文本中的语义信息融入文本的建模过程中，探讨其内在的逻辑和最终的情感倾向，形成细粒度的微博情感分类模型。本文设计了一种基于OCC模型和LSTM模型的财经微博文本情感分类的深度学习模型。其中，基于OCC模型建立认知情感角度建立情感规则，对财经微博文本进行情感分类标注作为训练集，通过word2vec生成的词向量作为输入，利用LSTM模型生成句子或篇章向量表示，接着将生成的句向量作为输入，调用TensorFlow平台的Keras模块完成对评论文本的情感分类，按照一定比例选取训练集进行训练，再将训练后的模型用于测试集测试，并对比训练模型所反馈的正确率，验证在不同模型下情感分类的正确率，从而证明模型的有效性。主要包括3个部分（图1）。

（1）输入层：给定一个待分类的财经微博文本，基于OCC模型，生成训练好的词向量集合，

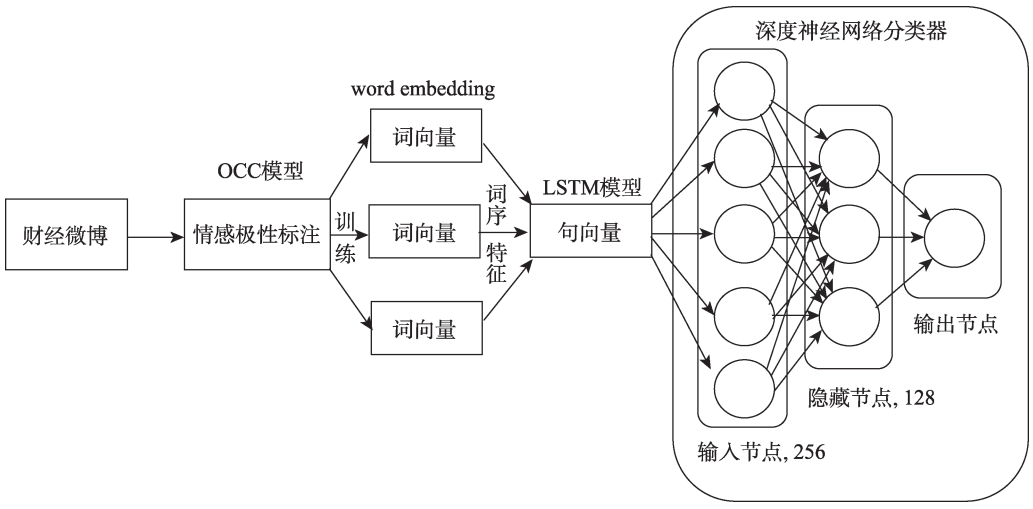

图1　财经微博情感分类处理流程

查找出各个词对应的词向量，生成输入矩阵。

（2）LSTM层：输入层生成的词向量矩阵经过LSTM网络，根据每个词对应的词向量时序性构建句向量，同时根据输入层生成的特征值序列获取输入文本中的词序信息，对于不同长度的文本，生成不同的特征值序列，并结合微博文本中有代表性的情感与语义信息，将每个特征值序列生成固定维度的特征。

（3）分类层：经过语义表达层的文本句向量输入TensorFlow平台Keras模块中，进行分类模型的构建，经过分类层后，输出两个情感类别（正面/负面）的概率，并将其和标准类进行比较，误差通过反向传播传递到模型前几层，然后更新其中参数。

3.1 财经微博情感规则和情感标注

本文基于OCC模型的逻辑规则框架和情感分类方式对微博进行标注，在OCC模型中，情感作为评价的认知过程的结果而产生的。产生过程可分解为以下四个过程[16]。

（1）分类：在OCC中评价取决于3种成分——事件的结果，关注情感分析的目标；对象的行为，关注个体行为准则；对象的描述，关注个体对事件的态度。

（2）量化：信息接受者接受信息的强度是否足以影响情感改变。

（3）映射：本文基于OCC模型将财经微博文本的情感映射为网民的正负面评价。

（4）表达：接受信息后产生的情感的表达形式，如面部表情、肢体行为以及文字等。

本文从事件结果和对象行为两个角度出发，根据事件演变的结果是否符合期望网民的期望目标，和事件中对象的行为是否符合网民的行为准则来判断财经微博文本的情感极性，模型如图2所示，情感识别函数定义如表1所示。

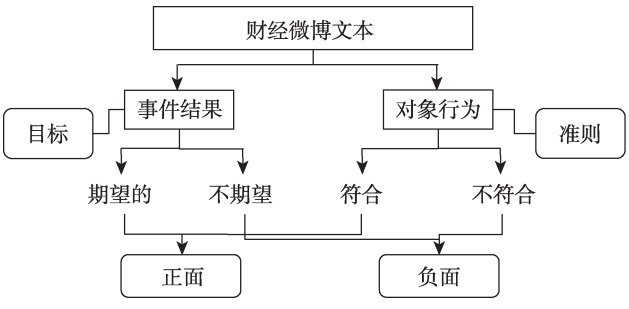

图2 基于OCC理论的网络舆情情感规则

表1 情感识别函数

函数	含义
EventOf(txt)	其值为句子所刻画的事件
EventConsequence(txt,e)	事件e的结果
ActionOf(e,l)	主体l在事件e中采取的行为
DesireOf(l,e)	主体l希望事件e发生
MotivationOf(e)	值为智能体做事件e的动机
PropertyOf(x)	其值为事件或对象的属性，值是一个集合
EffectOf(e)	其值为事件e产生的效果
ObiectOf(txt)	表示文本描述的对象
Focus(e)	其值为于事件相关的智能体
PolarityOf(l)	其值表示l的情感倾向
Agent(l)	若值为真，表示l是第一人称；若值为假，表示l是其他人称

上述情感识别函数可以用于识别OCC模型的逻辑规则框架，通过组合不同的情感识别函数，达到对微博进行分类和情感标注的目的。以上市公司“万达信息”的微博文本“昨天涨停卖了，今天5点又买进了大赚！”为例说明如何进行微博分类和标注。

（1）EventConsequence(txt,e)：从文本中得到的事件的结果是该网友“大赚”；

（2）ActionOf(e,l_1)：该网友采取了“涨停卖了”和“买进”的行为；

（3）DesireOf(l_1,e)：该网友期望得到大赚的结果；

（4）Agent(l)：该文本表示的主体是发表该文本信息的网友；

（5）Focus(e)：该网民关注的事件相关对象是自己。

由上述流程可以看出，该网民通过在股票涨停的时候抛售了股票并在次日又重新买进股票得到了大赚的结果，与其预期希望股票大赚相一致，从而将该微博文本信息分类为股民的微博信息，并由OCC模型的逻辑规则框架可以推理得出该情感应为正面情感，计算公式为

$$\text{EventConsequence}(\text{txt},e)\wedge\text{ActionOf}(e,l_1)\wedge\text{DesireOf}(l_1,e)\rightarrow\text{PolarityOf}(l)=\text{Positvie} \quad (1)$$

按照上述方式，本文对微博数据进行分类和标注，分别如表2和表3所示。如表中所示，当情感函数的取值为本人时，该微博属于股民和一般关注者的类别；当情感函数的取值为他人时，该微博则属于相关研究机构或此人。根据公式(1)得到的情感函数PolarityOf(l)的取值为positive时，微博的情感倾向性为正面；当情感函数PolarityOf(l)的取值为

negative时，微博情感倾向性为负面。

表2　微博分类

函数	取值	分类
Focus(e) ∨ Agent(l)	本人	股民和一般关注者
Focus(e) ∨ Agent(l)	他人	相关研究机构或个人

表3　情感标注

函数	取值	倾向性
PolarityOf(l)	positive	正面
PolarityOf(l)	negative	负面

3.2　情感倾向性分析

LSTM结构可以解决在复杂语言场景中，循环神经网络性能受到限制的问题。LSTM是一种特殊的循环体结构，不同于一般的循环体结构，它拥有三个“门”结构。所谓的“门”结构就是一个使用sigmoid神经网络和一个按位做乘法的操作，这两个操作合在一起就是一个“门”的结构。之所以该结构叫做“门”是因为使用sigmoid作为激活函数的全连接神经网络层会输出一个0~1的数值，描述当前输入有多少信息量可以通过这个结构。其中，文本信息特征提取包括以下内容。

（1）原始文本：根据总体框架设计，基于OCC模型的逻辑规则框架和分类方式，财经微博进行分类和标注后具有情感倾向性的文本。

（2）词向量：利用word2vec对原始文本进行分词处理，转化成word embedding高维向量。在传统的自然语言处理研究中，文本数据需要用one-hot方式表示出来，也就是用向量来表示每一个词，从而将一句话甚至一篇文档表示成一个高维向量。在每一个向量中，除了用于表示的那个词的位置是1之外，其余位置全部都用0表示。下文用买入和卖出两个词为例来形象地解释词向量的方式。

“买入”可以表示为：[1 0 0 0 0 0 0 0 0 0 0 0 …… 0 0 0 0 0]

“卖出”可以表示为：[0 1 0 0 0 0 0 0 0 0 0 0 …… 0 0 0 0 0]

利用这种方式，传统的机器学习方法包括朴素贝叶斯、支持向量机（support vector machine，SVM）、马尔科夫链、条件随机场等，都可以对自然语言进行处理。然而这种处理方式有一个弊端，以中文为例，词语的数量多达数十万，在一个one-hot模型中，如果数据样本足够大，每一个词语的编号范围则会从1到几十万，难以保证模型的稳定性。如果采用高维词向量，则可以解决多方向的发散问题。以20维向量为例，每一个维度仅仅需要0和1两个数字就可以表示100万左右的词语，小范围的变化可以提高模型的鲁棒性。本文基于word2vec来构造词向量（word embedding）。word2vec通过把相关度高的词语放在接近的位置，并使用实数向量来代替整数向量，从而获得对大量语料库的支持。

（3）句向量：根据LSTM模型将词向量构成的矩阵转化成低维的句向量。

（4）深度神经网络分类器：输入节点，本文所使用的深度神经网络输入节点的缺省参数为256个；隐藏节点，本文所使用的深度神经网络使用一层隐藏层，隐藏节点的缺省参数为128个。

（5）输出节点：经过深度神经网络分类器进行情感分类后，每一个句子都会得到一个基于LSTM框架模型计算后的情感倾向性。若与原始文本的情感倾向性相同，则为正确分类结果；否则，为错误分类结果。

4　实证研究

4.1　实验设计

本文的实验目的如下。

（1）构建基于OCC情感规则和LSTM模型的财经微博文本情感识别模型（OCC-LSTM模型）。

（2）验证和对比分析本文构建的OCC-LSTM模型的科学性。

在上述模型构建的基础上，本文设计实验流程如图3所示，使用TensorFlow平台Keras模块完成实验研究内容，使用Nvidia公司的GPU来辅助运算。

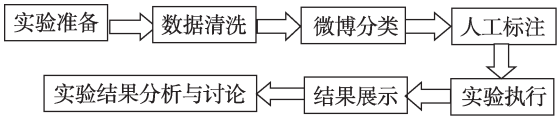

图3　实验流程设计

4.2　数据处理

本实验从微博上采集了13家上市公司2014年1月—2017年5月的微博舆情信息，如表4所示。数据清洗去除原始数据中的噪声数据、无关数据、遗漏数据和清洗脏数据、空缺值等[24]。本文处理方法如下。

（1）噪声：噪声数据是指被测量的变量的随机误差，在本文研究中指和上市公司无关的微博舆情信息，该类噪声数据在微博评论中存在数量较高，

表 4 上市公司微博文本内容示例

微博 id	内容	公司名称	发布时间
1001446372_CzrelExok	今日盘面,互联网软件服务现涨停潮,万达信息、东方国信、柘尔思、卫宁软件等 18 只个股涨停。工程机械、电脑设备等涨幅居前。银行板块有补跌嫌疑拖累大盘,交通银行跌停,工行跌 7.8%,建行跌 7.8%。	万达信息	2015-09-07,15:34
1002344165_E6ZXOrAFV	中亚股份、贵人鸟、万达信息、美年健康、恒生电子、东南网架等一大批好票都进入射程范围内了,现在就缺指数的一个急跌来配合下。	万达信息	2016-09-06,12:09
1003722482_DcQ0Zdfm5	我推荐的中国平安虽显得不温不火,但大跌过程很抗跌反弹的时候不落后说明有大资金介入,昨天最重要是新医保制度这将影响每个人的生活,推荐万达信息 300168、卫宁健康 300253,另外我会推荐相关基金长线布局该产业,仅供参考。	万达信息	2016-01-12,23:07

需要清洗。

(2) 空缺值:一些微博数据因为内容过长,抓取数据时只能抓取到“显示全文”的关键词,无法抓取到全部数据,影响模型的训练效果,需要清洗掉。

(3) 脏数据:微博的转发数据中存在大量内容只有“转发微博”关键词数据,这部分数据也无法对微博情感分类模型的训练效果有提升作用,因此也需要删除。除此之外还有一些结构不完整或者重复的数据也属于脏数据需要清洗掉。

经过清洗之后的上市公司微博舆情信息如表 5 所示。

表 5 上市公司微博舆情信息

公司名	数据量
国脉科技	1499
捷成股份	1402
南天信息	668
启明信息	4261
四方精创	1903
四维图新	7487
同有科技	1431
卫宁健康	1462
卫士通	4785
佳都科技	1813
世纪瑞尔	638
万达信息	6185
中海科技	1571

4.3 基于 OCC 模型的微博分类和人工标注

本文基于 OCC 模型对财经微博数据进行分类,根据图 1 处理流程,以万达信息为例,剔除官方的微博数据后,保留相关研究机构和股民的微博数据,如表 6 和表 7 所示。如表 7 所示,用上述的方式,将 13 家企业的微博舆情数据都划分成两类,为后续对财经领域微博数据的情感分类研究提供训练集和测试集数据的准备(表 8)。因为 RNN 和 LSTM 结构都是属于有监督的机器学习内容,所以需要通过人工标注的方式来设计一定数量的训练集数据来训练本文的深度学习情感分类模型。人工标注方式参照图 2 介绍的 OCC 模型的逻辑规则框架和图 1 的设计框架,分别对两类微博数据进行正负面的情感倾向性标注。

4.4 基于深度神经网络的情感分类实验

根据图 1 构建的 LSTM 模型,使用深度学习中的 TensorFlow 框架和 Keras 模块建立相应的实验模型,建立 Keras 的序列化模型,设定 LSTM 模型各项参数,包括训练轮数 epoch 取值实验、Dropout 取值实验,对上述处理好的微博数据进行分析。实验流程如图 4 所示,经过训练后得到的算法结果、损失值和准确率随训练轮数的变化如图 5 所示。从图 5 可以看出,在训练次数到达 10 轮之后,训练集的准确率已经逐渐接近 1;在训练次数到达 26 轮之后,信息损失值已经逐渐接近 0 了。说明该实验模型具有一定的实用价值。

表 6 相关研究机构和个人微博舆情信息

微博 id	内容	来源	发布时间
1235206304_BtwSK31EA	此项目中标,使万达信息不仅获得了首个省级规模医药供应链网络商务平台 建设的机会,而且标志着公司正式进入药品服务领域,也在国家三医联动(“医疗—医药—医保”)的行业改革方向上,走向更深、更广的市场……	万达信息	2014-10-27,20:11
1235206304_BtwSK31EA	万达信息、卫宁软件、海虹控股这几个今天都新高了,一个板块的医疗服务类股票已经连续牛了好几月。	万达信息	2014/10/27,20:11
1365253073_C6Ow0toyg	万达信息城市信息化细分行业优势明显,并购推进扩张_财经频道_东方财富网(Eastmoney.com)|万达信息城……	万达信息	2015-03-03,09:32

表 7　股民和其他领域财经新闻关注者舆情信息

微博id	内容	来源	发布时间
1296527963_DvYxsn5sO	一个午觉醒来，我的万达信息就封停啦！	万达信息	2016-05-17，19：45
1642585887_BwIVnCTg0	中国经济已经步入故事时代。但傻瓜多，天还未亮，所以股市还会涨。//@晓峰_徐：这个带几万亿，哪个区几万亿，今年以来，从渤海湾到长三角到珠三角，钱在哪里都不知道，一路画饼，画完就搁着，股市炒完就扔。太搞笑了！//@止损之剑：又开始忽悠散户了。	万达信息	2014-11-17，22：32

表 8　微博舆情数据分类结果

类别	数据量
相关研究机构或个人发表微博	21107
股民及其他财经领域新闻关注者发表微博	12896

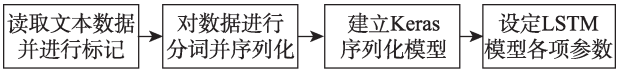

图 4　算法流程图

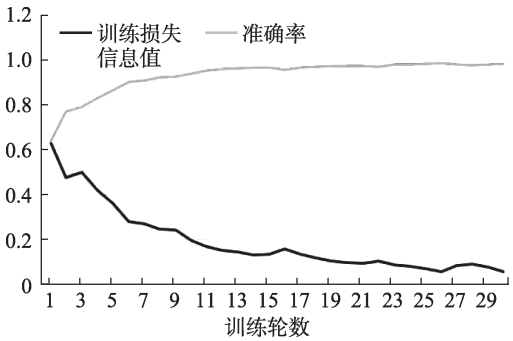

图 5　训练集训练结果

4.5　实验结果讨论

通过上述的实验内容，得到了表 9 所示的实验结果，实验结果表明，基于OCC模型的财经领域微博网络舆情情感分类研究，在相关研究机构或个人发表微博中，训练准确度可以达到98.39%，测试准确度也达到了89.03%，超过了目前大部分基于机器学习的情感分类研究80%左右的结果。而在股民及其他财经领域新闻关注者的发表微博中，训练准确度达到了95.59%，测试准确度可以达到89.03%。从结果可以得出以下结论。

（1）基于OCC模型对财经领域微博情感分类建模研究，将网民和舆情类型进一步划分之后，可以得到较高的准确度，从而证明了OCC模型具有较高的实用价值。

表 9　两类微博实验结果数据

类别	训练轮数	训练准确度/%	测试准确度/%
相关研究机构或个人发表微博	30	98.39	89.45
股民及其他财经领域新闻关注者发表微博	30	95.59	89.03

（2）针对财经领域的网络舆情进行情感分类研究，在模型中加入财经领域特有的情感词，可以提高情感分类的准确度，从而也证明了目前网络舆情情感分类研究需要和一定的领域相结合。

（3）相关研究机构或个人发表的微博内容，相较于股民及其他财经领域新闻关注者的微博内容，具有更高的训练准确度和测试准确度，证明研究机构或个人的微博内容具有更加明确的情感倾向特征。股民及其他财经领域新闻关注者的微博信息中表达情感观点的方式更加复杂。

4.5.1　LSTM计算参数检验

1）训练轮数epoch取值实验

在循环神经网络中，每一轮epoch包括所有的训练操作[25]，虽然理论上训练的轮数越多会越有利于模型用于测试的准确率的提高，但是当训练轮数到达一定的轮数时，由于存在过拟合的问题，反而会导致准确率下降，同时也会增加计算机运算的负担，使得模型的实用性大大降低。因此本文通过改变epoch次数来探究一个合理的训练次数。

实验过程中，选取股民及其他财经领域新闻关注者发表的微博进行epoch次数选取实验。实验结果如图6所示。

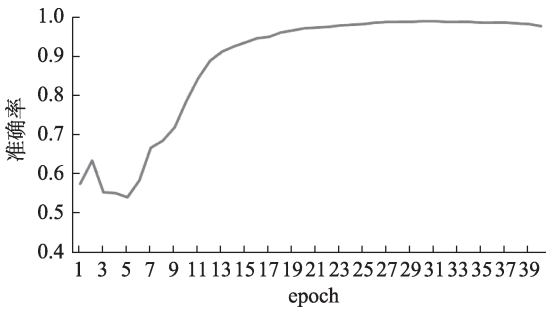

图 6　epoch选取测试准确率实验结果

从图6可以看出，epoch从1增加到15的过程中，测试准确率有着明显的提升；在从15增加到30的过程中，测试准确率在缓慢增加，当epoch取

到30时，测试准确率达到最大值；当epoch取值继续增加时，测试准确率出现下降的趋势。证明epoch选值30是最适合本实验模型的取值。

2）Dropout取值实验

大部分的机器学习分类研究领域都要考虑过拟合的现象。过拟合是指在使用一个统计模型中，由于使用了过量的参数导致假设的过度拟合，在训练集取得较好的结果但在测试集中难以获得较高的准确率的问题。

Dropout技术[25]可以通过在训练过程中随机地舍弃前面一层中某些神经元中的信息来减少过拟合现象，但是过度使用Dropout技术则会导致有效信息被过量抛弃，影响模型的最终准确性。本节通过对不同的Dropout取值进行测试来寻找一个较为理想的参数。以第一类微博的情感分类模型为例，实验结果如表10所示。

表10 Dropout取值实验结果

Dropout	训练准确度/%	测试准确度/%
0	99.39	88.63
0.1	99.21	88.95
0.2	98.84	88.99
0.3	98.92	89.34
0.4	98.44	89.39
0.5	98.39	89.45
0.6	98.28	89.33

从表10可以看出，随着Dropout取值不断增加，训练准确度不断下降，过拟合的问题得到了解决。当Dropout取值从0增加到0.5的时候，测试准确度不断提高；但当Dropout增加到0.6的时候，测试准确度有所下降，说明Dropout技术使用过度，导致有效信息丢失，影响了实验研究的准确性。因此，对Dropout选取0.5是经过实验验证的最适合本文实验研究的取值。

4.5.2 结果对比分析

近几年对微博网络舆情的情感分类研究方法主要是SVM和深度学习，本文根据朱少杰等[25]的研究结果和其对比研究进行实验结果分析，分析结果如表11所示。

表11 结果对比分析

研究方法	准确度/%
基于SVM方法的多特征组合情感分类研究	81.88
基于深度学习的半监督RAE方法的情感分类研究	85.10
基于OCC-LSTM模型的情感分类研究	89.45

通过对比可知，本文基于OCC模型和深度学习，针对财经领域微博网络舆情进行情感分类研究，获得了较高的准确率。该实验结果一方面由于本文基于OCC模型构建了情感分类模型，对不同网民类型的微博信息进行了分类，降低了机器学习的难度，另一方面由于于财经领域的情感词都有较为明显的特征，便于情感模型的构建和机器学习的特征提取和学习。

5 结 论

本文基于OCC模型和LSTM模型，建立了财经微博文本情感分类OCC-LSTM模型，结合财经舆情领域的13家上市公司近3年来的网络舆情数据，考虑了不同网民个体在同一财经舆情事件下的情感观点差异，在TensorFlow平台下使用LSTM模型进行了情感分类的建模和实证研究。最终实证研究结果取得了89.45%的准确率，高于同类型的研究的准确率，证明了OCC模型和深度学习在财经领域微博网络舆情的情感分类研究中具有一定的科学性和先进性，为相关研究机构在进行分析和决策的时候提供了舆情信息的支持依据，也为网络舆情情感分类研究提供了一个新的解决方法。

目前国内对于社交媒体特别是微博情感分类的研究日趋丰富，但基本上都是从文本本身的角度出发，很少结合认知心理学的相关理论开展研究，忽略了不同网民个体在发表微博新闻和评论时带有的情感观点的差异，没有从认知情感评价的角度对微博情感进行分析研究。本文设计的OCC-LSTM模型，从认知情感角度设计OCC情感规则标注训练集来提高训练集的数据准确性，采用LSTM模型，不使用情感词典和句法分析结果，仅使用少量标注的训练集和测试集，对财经微博文本进行情感分类，考虑了不同网民个体在对于同一个新闻事件下不同的情感观点差异，从而提出一种情感分类的新的研究方法。

参 考 文 献

[1] Kennedy A, Inkpen D. Sentiment classification of movie reviews using contextual valence shifters[J]. Computational Intelligence, 2006, 22(2): 110-125.

[2] 滕飞, 郑超美, 李文. 基于长短期记忆多维主题情感倾向性分析模型[J]. 计算机应用, 2016, 36(8): 2252-2256.

[3] Piñeiro-Chousa J R, López-Cabarcos M Á, Pérez-Pico A M. Examining the influence of stock market variables on microblog-

ging sentiment[J]. Journal of Business Research, 2016, 69(6): 2087-2092.

[4] Shen D H, Liu L B, Zhang Y J. Quantifying the cross-sectional relationship between online sentiment and the skewness of stock returns[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 490: 928-934.

[5] Ruan Y F, Durresi A, Alfantoukh L. Using Twitter trust network for stock market analysis[J]. Knowledge-Based Systems, 2018, 145: 207-218.

[6] Chen M Y, Chen T H. Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena[J]. Future Generation Computer Systems, 2019, 96: 692-699.

[7] 梁军, 柴玉梅, 原慧斌, 等. 基于极性转移和LSTM递归网络的情感分析[J]. 中文信息学报, 2015, 29(5): 152-159.

[8] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[9] Ortony A, Clore G L, Collins A. The cognitive structure of emotions: Factors affecting the intensity of emotions[J]. Contemporary Sociology, 1988, 18(6): 2147-2153.

[10] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014, 2: 3104-3112.

[11] Li J W, Luong M T, Jurafsky D. A hierarchical neural autoencoder for paragraphs and documents[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1106-1115.

[12] Nguyen N K, Le A C, Pham H T. Deep bi-directional long short-term memory neural networks for sentiment analysis of social data[C]// Proceedings of the 5th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making. Cham: Springer, 2016: 255-268.

[13] Wang X, Liu Y C, Sun C J, et al. Predicting polarities of tweets by composing word embeddings with long short-term memory [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2015: 1343-1353.

[14] Cheng J J, Zhang X, Li P, et al. Exploring sentiment parsing of microblogging texts for opinion polling on chinese public figures [J]. Applied Intelligence, 2016, 45(2): 429-442.

[15] 陈鹏. 基于深度语义特征的情感分析研究[D]. 哈尔滨: 哈尔滨工业大学, 2016.

[16] Adam C, Herzig A, Longin D. A logical formalization of the OCC theory of emotions[J]. Synthese, 2009, 168(2): 201-248.

[17] Clore G L, Palmer J. Affective guidance of intelligent agents: How emotion controls cognition[J]. Cognitive Systems Research, 2009, 10(1): 21-30.

[18] Jaques P A, Vicari R M. A BDI approach to infer student's emotions in an intelligent learning environment[J]. Computers & Education, 2007, 49(2): 360-384.

[19] Bartneck C. Integrating the OCC model of emotions in embodied characters[C]// Proceedings of the Workshop on Virtual Conversational Characters: Applications, Methods, and Research Challenges. Melbourne, 2002.

[20] Roberts K, Roach M A, Johnson J, et al. EmpaTweet: Annotating and detecting emotions on Twitter[C]// Proceedings of the Eighth International Conference on Language Resources and Evaluation, 2012: 3806-3813.

[21] Yalcin Ö N, DiPaola S. A computational model of empathy for interactive agents[J]. Biologically Inspired Cognitive Architectures, 2018, 26: 20-25.

[22] Deng J J, Leung C H, Mengoni P, et al. Emotion recognition from human behaviors using attention model[C]// Proceedings of the First International Conference on Artificial Intelligence and Knowledge Engineering. New York: IEEE Computer Society, 2018: 249-253.

[23] Deng J, Leung C, Li Y X. Beyond big data of human behaviors: Modeling human behaviors and deep emotions[C]// Proceedings of the Conference on Multimedia Information Processing and Retrieval. New York: IEEE Computer Society, 2018: 282-286.

[24] 梁军, 柴玉梅, 原慧斌, 等. 基于深度学习的微博情感分析[J]. 中文信息学报, 2014, 28(5): 155-161.

[25] 朱少杰. 基于深度学习的文本情感分类研究[D]. 哈尔滨: 哈尔滨工业大学, 2014.

（责任编辑　王克平）