

## 一种基于深度学习的中文微博情感分析融合模型

周映宇

(北京邮电大学理学院, 北京 100876)

**摘要:**随着网络信息技术的不断发展,文本挖掘与处理工作渐渐交由电脑执行,而情感分析便是这样的一种技术。作为一种自然语言处理方法,情感分析能够有效的分析出文本信息中所包含的主观意愿,在商品满意度调查、微博态度判别等方面都有着广阔的应用。本文首先对前人在情感分析方面的工作做了梳理,并针对传统的深度学习特征提取不充分的问题,提出了一种融合的深度学习模型。该模型首先通过双向长短期记忆网络(BiLSTM)提取文本的长距离时序特征,然后以此为输入,利用卷积神经网络(CNN)提取空间特征。进一步,利用一种双层注意力结构提取原文本的时序特征。最后将这两部分提取到的特征进行拼接送入全连接层,然后进行分类。实验结果表明:本文提出的模型相较于单独的传统模型和部分混合模型的准确度与F1值都有所提升,证明了本文所提出模型的有效性。

**关键词:**深度学习;文本情感分析;卷积神经网络(CNN);双向长短期记忆网络(BiLSTM);注意力机制  
**中图分类号:**TP181      **文献标识码:**A      **文章编号:**1004-7344(2022)36-0151-03

## 0 引言

随着人们互联网的参与度的逐年提高,网络信息的体量也呈指数级增长,面对如今的网络信息洪流,传统的人工方法很难在庞大的信息海洋中找到有用的信息。因此,如何高效的挖掘和利用文本形式的交互资源中所包含的信息,成为目前计算机领域的热门研究方向。而最近兴起的文本情感分析相关技术正是解决此问题的方法之一。现有的情感分析技术大体上分为三类:基于统计的、基于机器学习的以及基于深度学习的情感分析。基于统计的情感分析方法<sup>[1]</sup>的主要模式是建立情感词库,赋予情感词语极性,然后通过情感词库对每个句子的语义及情感极性做出计算。但是这类方法过于依赖情感词库,且情感词库需要不断更新,模型的泛化能力较弱。因此,机器学习的相关方法被引入了情感分析领域中。基于机器学习的情感分析方法<sup>[2-3]</sup>的主要模式是通过人工对文本进行分析,使用相应的模型来构造文本的特征表示,然后输入分类器(如支持向量机 SVM)中进行分类。但是人工构建文本特征需要一定的先验知识(比如 LDA 主题模型等等),因此提取到的特征信息相对不完善。基于深度学习的情感分析方法<sup>[4-5]</sup>不再需要人工来构造文本特征,通过深层次的神经网络,模型可以自行对每个文本构建不同方面的特征信息用于分类。但不同的深度学习模型提取的特征有不

同的侧重(如 CNN 难以提取文本长依赖特征),导致同一个模型在不同的数据集上的效果可能并不相同(如卷积神经网络 CNN 在短文本上的效果一般比长文本好<sup>[6]</sup>)。对于上述问题,本文根据已有研究,将卷积神经网络(CNN)、双向长短期记忆网络(BiLSTM)、注意力机制等模型的特点进行了充分的融合,提出了一个基于深度学习的中文微博情感分析融合模型。

## 1 本文模型的构建

## 1.1 模型的整体结构

本文模型的整体结构如图 1 所示。主要流程分为文本预处理、文本词嵌入、特征提取、情感分类 4 个步骤。

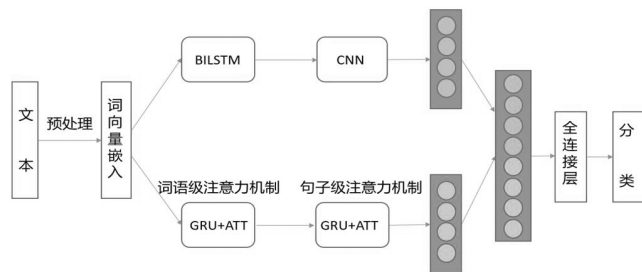


图 1 本文模型总体结构

图 1 中,模型首先对文本进行了预处理,将中文句子拆分为单独的词语并去掉了无意义的标点符号及停用词。然后对每个词进行词嵌入,这里使用的是 Word2vec。接着通过 BiLSTM 提取文本的长距离时序特征,然后以此为 CNN 的输入来提取微博的空间特征。

与此同时,利用一种双层注意力结构提取原文本的时序特征。最后将这两部分提取到的特征进行拼接,以全连接层相接并进行分类。

## 1.2 文本预处理

深度学习模型不能直接对中文微博的原文本进行处理。在将数据输入模型处理之前,需要对数据进行预处理。主要完成了以下3个步骤:①过滤无实际意义的标点等符号,只保留包含语义信息的文本。②使用jieba对剩下的中文文本进行分词,使每一个句子都变成一个词语的形式。③去掉一些对情感分析无意义的停用词(如虚词等)。

## 1.3 文本词嵌入

在预处理流程过后,文本变成了拥有实际意义的词语的集合。但是即使分成了词语的集合,还需要将每一个词语转换为计算机能够识别的语言,即词语的向量化。这个过程便被称为词嵌入。本文使用的是Word2vec词袋模型,通过Word2vec算法计算出每个词语相对应的词向量。在训练完成之后,我们不仅可以得到不同单词相对应的词向量,还可以实现诸如寻找某一个词的最相近的词集合、比较两个词的相似程度等功能。

## 1.4 特征提取

在将文本中的词语向量化后,就可以使用模型进行特征提取了,本文主要使用的是CNN、BiLSTM、注意力机制等模型。本文模型的特征提取模块分为上下两部分。上半部分为了将BiLSTM提取的时序特征和CNN提取的空间特征结合起来,采用了一个BiLSTM和CNN的串联结构。通过实验,串联结构是BiLSTM和CNN效果最好的结合方式。下半部分为了防止在句子中出现的较早的情感词所包含的信息丢失,采取了一个注意力结构来动态为每个词语和句子分配权重。最后,两部分得到的特征做一个拼接,作为整体模型提取的特征。

### 1.4.1 BiLSTM和CNN的串联模块

卷积神经网络(CNN)是一种最为经典的神经网络,由于卷积核的存在,能够有效捕获文本的局部空间信息。与图形处理CNN不同的是,针对文本处理的卷积神经网络的卷积核并不是正方形的,且滑动窗口大小一般远远小于词嵌入的维度,因此卷积核为长方形。

STM是RNN的一种,能够有效的提取文本的时序

特征,并且解决了传统循环神经网络的长期依赖和梯度消失问题,但其只能关注每个单词的上文信息。BiLSTM在前者的基础上又做出了改进,使得模型能够同时关注到词语的上下文信息。一般使用BiLSTM的隐藏层向量作为文本的特征向量进行分类,这里由于该模型的输出与卷积神经网络CNN的输入比较类似,可以在后面接入一个CNN网络来继续提取文本的空间特征。

在BiLSTM和CNN的串联模块中。首先是输入层:这里完成了文本的预处理及词嵌入,使用Word2vec词袋模型,词向量大小设置为100维。接下来使用BiLSTM提取文本的时序特征。进入BiLSTM的数据的维度为[batch\_size, seq\_len, embedding\_dim],BiLSTM的输出维度是[batch\_size, seq\_len, hidden\_dim\*2]。这里batch\_size为批样本个数,seq\_len为文本长度,embedding\_dim为词向量维度,hidden\_dim为BiLSTM隐藏层节点个数,这里取256。之后使用CNN提取文本空间特征。卷积核大小为3\*512、4\*512、5\*512,每种卷积核使用100个。CNN的输入维度为[batch\_size, 1, seq\_len, hidden\_dim\*2],这里相比BiLSTM的输出多了一个维度,代表通道数为1。经过CNN处理后,通过一个全连接层将不同卷积核得到的输出进行拼接,作为此模块的总体输出,输出维度为[batch\_size, 3\*100],这里的3对应卷积核的种类,100对应每个种类卷积核的个数。

### 1.4.2 注意力机制模块

注意力机制是一个模仿人类注意力的机制,在循环神经网络(RNN)中,随着句子序列长度的增加,越靠前的词语所包含的信息越容易丢失。为此,注意力机制使用了一个上下文向量来动态的为每个词语分配权重,以保证重要的词语(如情感词)所包含的信息不会丢失。

这里将每一条文本看作若干个句子组成,每一个句子看作若干个词语构成。进行情感分析时,不同的词语对每个句子的重要度不一样,不同的句子对每个文本的重要程度也不一样。因此,这里采用了词语级注意力机制和句子级注意力机制的两层结构。由于两层结构类似,这里对词语级注意力机制进行一个详细说明。首先是一个词语级编码器,这里使用门控循环网络GRU(LSTM的一种较为简单的变体)来进行编码。设句子i含有T个词语 $w_{it}$ , $t \in [1, T]$ ,则每个词语的表征向量

$h_{it}$  可以由式(1)-式(4)得到。

$$x_{it} = W_e w_{it}, t \in [1, T]. \quad (1)$$

$$\vec{h}_{it} = \overrightarrow{\text{GRU}}(x_{it}), t \in [1, T]. \quad (2)$$

$$\overleftarrow{h}_{it} = \overleftarrow{\text{GRU}}(x_{it}), t \in [T, 1]. \quad (3)$$

$$h_{it} = [\vec{h}_{it}, \overleftarrow{h}_{it}]. \quad (4)$$

其中:  $W_e$ ——词嵌入矩阵, 接下来经过词语级注意力层得到句子  $i$  的表征向量  $s_i$ , 具体计算过程如下:

$$u_{it} = \tanh(W_w h_{it} + b_w). \quad (5)$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)}. \quad (6)$$

$$s_i = \sum_t \alpha_{it} h_{it}. \quad (7)$$

其中:  $W_w, b_w$ ——注意力层的权重矩阵和偏置;  $u_w$ ——上下文向量, 在训练过程中被随机初始化和联合学习。

### 1.5 情感分类

经过上面的步骤, 已经得到了两个单独的文本表征, 接下来将两种途径获得的文本表征进行拼接, 通过全连接层接入后进行分类。分类函数选用 sigmoid。本文的损失函数为二元交叉熵函数, 学习率采用  $10^{-3}$ 。

## 2 实验与结果分析

### 2.1 实验数据集

本文采用的情感数据集是一个开源数据集, 一共有 12 万条带情感标注的新浪微博评论, 其中正负向评论数量相同, 是一个较为平衡的数据集。数据集中取八成作为训练集。

### 2.2 评价标准

混淆矩阵是一个评判分类器效果的常用方法。参如表 1 所示。本文所关注的主要评价标准为准确率和 F1 值。准确率的计算公式为  $A = (TP + TN) / (TP + FP + FN + TN)$ 。F1 值的计算公式为  $F1 = 2RP / (R + P)$ 。其中, 准确率代表了模型分类正确的样本数占总样本数的比率。F1 值主要作用是防止出现样本中 90% 都为正样本, 而模型将所有样本判定为正即可达到 90% 准确率的无效模型, 是一个综合评价指标。

表 1 混淆矩阵

	正类(真实)	负类
正类(预测)	TP	FP
负类(预测)	FN	TN

### 2.3 模型结果比较

表 2 为模型结果比较。

表 2 本文模型与其他模型比较

模型	LOSS	准确率/%	F1/%
CNN	0.116	96.48	96.38
BiLSTM	0.101	96.46	95.91
BiLSTM+CNN	0.080	97.34	96.80
GRU+ATT	0.079	97.44	97.17
本文方法	0.068	97.85	97.33

表 2 展示了本文提出的模型与其他模型的效果对比。其中 BiLSTM+CNN 是串行方法, GRU+ATT 代表上面提到的注意力模块。相较于单独的模型, 融合模型无论是在准确率, 还是在 F1 值上面都有不错的提升, 这说明单独的 CNN 或 BiLSTM 在提取语义特征信息时是有一定缺失的。另外本文的模型相较于自身两个单独的模块在效果上也有一定的提升, 这说明本文的模型结合方式也是确实有效的。

## 3 结语

本文针对传统深度学习模型的不足, 给出了一个基于深度学习的情感分析融合模型。实验表明: 本文提出的模型相较于单独的传统模型和部分混合模型的准确度与 F1 值都有所提升, 证明了本文所提出模型的有效性。

### 参考文献

- [1] 王科, 夏睿. 情感词典自动构建方法综述[J]. 自动化学报, 2016, 42(4): 495-511.
- [2] Pang B, Lee L, Vaithnathan S. Thumbs up? Sentiment classification using machine learning techniques[C]// Proceedings of the conference on Empirical Method in Natural Language Processing. Philadelphia, 2002: 79-86.
- [3] 何跃, 邓唯茹, 张丹. 中文微博的情绪识别与分类研究[J]. 情报杂志, 2014, 33(2): 136-139, 85.
- [4] Poria S, Cambria E, Gelbukh A. Deep Convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 2539-2544.
- [5] 杨艳, 徐冰, 杨沐昀, 等. 一种基于联合深度学习模型的情感分类方法[J]. 山东大学学报(理学版), 2017, 52(9): 19-25.

作者简介: 周映宇(1996—), 男, 汉族, 重庆人, 硕士研究生, 研究方向为自然语言处理。