**Modeling and Simulation** 建模与仿真, 2022, 11(5), 1248-1254
Published Online September 2022 in Hans. http://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2022.115116

Hans汉斯

# 基于LSTM的文本情感分析方法

**郭　霖，高允轩**

湖北工业大学计算机学院，湖北　武汉

收稿日期：2022年7月15日；录用日期：2022年8月26日；发布日期：2022年9月6日

## 摘　要

**随着互联网的发展和电子商务的兴起，人们通过各种社交、电商平台发表自己的看法与见解，从这些用户评论数据中准确地挖掘出有用的信息是当前的研究热点。针对网络中的各类文本评论数据，本文基于深度学习的方法对这些数据进行情感分析，采用长短期记忆(Long Short-Term Memory, LSTM)神经网络模型构建情感分类器，对文本的情感倾向进行预测与分类。实验表明基于LSTM的情感分析方法可以很好地解决长距离依赖问题，具有较好的分类效果。**

## 关键词

**情感分析，长短期记忆，神经网络，深度学习**

# Text Sentiment Analysis Method Based on LSTM

**Lin Guo, Yunxuan Gao**

School of Computer Science and Technology, Hubei University of Technology, Wuhan Hubei

Received: Jul. 15th, 2022; accepted: Aug. 26th, 2022; published: Sep. 6th, 2022

## Abstract

**With the development of the Internet and the rise of e-commerce, people express their views and opinions through various social and e-commerce platforms. It is a challenging issue to accurately mine useful information from user comment data. For all kinds of text comment data in the network, sentiment of the comment data is analyzed in this paper through deep learning. The Long Short-Term Memory (LSTM) neural network model is used to construct a sentiment classifier to predict and classify the sentiment tendency of the text. The experimental results show that the sentiment analysis method based on LSTM can solve the problem of long-distance dependence and has a good classification effect.**

文章引用: 郭霖, 高允轩. 基于 LSTM 的文本情感分析方法[J]. 建模与仿真, 2022, 11(5): 1248-254.
DOI: 10.12677/mos.2022.115116

**Keywords**

**Sentiment Analysis, Long Short-Term Memory, Neural Network, Deep Learning**



## 1. 引言

近年来，在网络的影响下，人们可以通过各大电商平台在网络上进行购物，并在使用后对商品进行评论，或在社交平台上表达自己对于某件事的看法、感想等。我们可以通过提取网络上的这些情感信息从而分析消费者的想法以及舆论的倾向等，为后续改进的方向与决策的实施提供便利。情感分析主要以计算机技术为基础，分析电商平台与社交平台评论的情感倾向[1] [2]，从而挖掘出更多重要信息。目前该项技术已经广泛应用于政治与经济等领域[3]。

本文利用深度学习方法构建了情感分类模型，提出基于长短期记忆(Long Short-Term Memory, LSTM)神经网络模型的情感分类方法，通过实验分析比较了本文方法与卷积神经网络(CNN)、循环神经网络(RNN)性能的差异，实验结果表明本文方法可以很好地解决长距离依赖问题，具有较好的分类效果。

## 2. 相关工作

情感分析的研究历程主要分为基于情感词典的方法、基于传统机器学习的方法、基于深度学习的方法三个阶段。

基于情感词典[4]的情感分类方法主要是将文本中的单词与情感词典中的词进行匹配从而得到文本的情感倾向。该方法的分类质量主要取决于情感词典是否全面且精确。然而，由于分类结果对情感词典的依赖性较高，网络时代新的词语与文本诞生的速度很快，使得该方法不能很好地应用于实时评论数据分析。

基于机器学习的情感分类算法[5] [6]是利用训练好的分类器对文本的情感倾向进行分类。比较常用的方法有朴素贝叶斯方法，主要通过计算先验概率、后验概率、条件概率等来对文本的情感进行分类，该方法对小规模数据表现良好；最大熵方法主要通过计算文本的熵值来进行分类；支持向量机方法通过核方法来进行优化，并加入了正则项来提高模型的泛化能力。与基于情感词典的方法相比，机器学习方法提高具有一定的自主性，不需要制作特定词典进行分类，只需要对标记好的语料库进行训练即可。但对语料进行标注需要花费大量人力与物力[7]。

随着网络的发展，传统的情感分析方法在处理文本数据时效率低下。而随着深度学习的出现与发展，基于深度学习的情感分析模型逐渐发展壮大。常用的基于深度学习的情感分析算法主要包括：卷积神经网络 CNN [8]、循环神经网络 RNN [9]以及由 RNN 改进而来的长短期记忆网络 LSTM [10]。其中，CNN 模型存在许多局限性，虽然其通过卷积层对特征图进行了提取，获取了文本中最重要的特征，但同时也失去了该特征的位置信息，而位置信息在文本情感分析中往往十分重要。RNN 的出现使文本序列问题有了解决方法，但还是无法解决文本的长距离依赖问题。而 LSTM 网络通过添加记忆单元使得网络能够处理长距离信息，适用于文本情感分析问题。因此，本文基于 LSTM 构建文本情感分类模型，对文本的情

感倾向进行预测与分类。

## 3. 基于 LSTM 的情感分类模型

LSTM 的主要原理是在隐藏层添加一种特殊的记忆单元来保存长距离信息，并加上了特殊的门结构，包括输入门、输出门和遗忘门，这些门结构会协助处理信息在网络中的传递，并且记忆单元的状态更新也依赖于三个门结构的控制[11]。图 1 为 LSTM 的单元结构图。

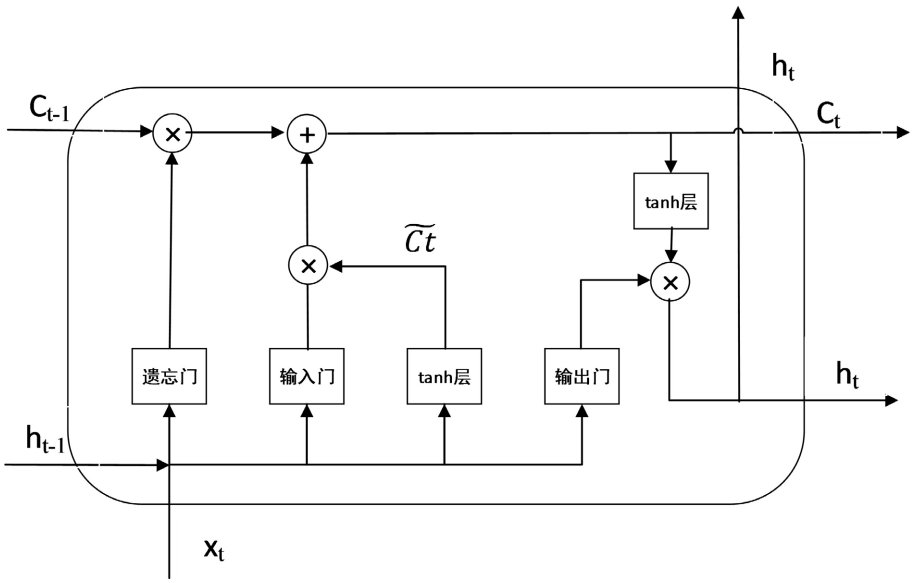


**Figure 1.** LSTM network structure
**图 1.** LSTM 网络结构图

LSTM 的关键在于记忆单元，首先该单元需要决定丢弃哪些无用的信息，这一步操作主要通过遗忘门来实现。遗忘门采用的激活函数 σ 为 sigmoid 函数，如公式(1)所示：

$$f_t = \sigma\left(W_f \cdot \left[x_t, h_{t-1}\right] + b_t\right) \tag{1}$$

其中，$W_f$ 表示遗忘门的权重矩阵，$b_t$ 表示遗忘门的偏置。获得的结果 $f_t$ 的值域为[0,1]，$f_t$ 通过 $t$ 时刻的输入 $x_t$ 以及 $t$-1 时刻的输出 $h_{t-1}$ 的内容来决定该单元需要丢弃的信息。当 $f_t$ 的取值为 0 时，该单元则会丢弃过往传递来的所有信息；当 $f_t$ 的取值为 1 时，该单元则会保留所有的信息；而当 $f_t$ 的取值为 0.5 时，该单元则会丢弃部分信息。相比于 RNN 网络始终保留所有信息的特点，遗忘门的操作使得 LSTM 网络能够控制依赖信息的取舍。

记忆单元经过遗忘门的处理后，第二步则是选择需要选择向信息传送带中加入的哪些新的信息，主要通过输入门来实现。添加的新信息主要分为两个部分，首先，采用 $\sigma$ 函数对两个输入信息 $x_t$ 和 $h_{t-1}$ 进行处理，根据公式(2)获取有用的新信息 $i_t$。其次，利用公式(3)计算需要加入的候选信息 $\tilde{C}_t$。

$$i_t = \sigma\left(W_i \cdot \left[x_t, h_{t-1}\right] + b_i\right) \tag{2}$$

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[x_t, h_{t-1}\right] + b_c\right) \tag{3}$$

其中，$W_i$ 和 $W_c$ 分别表示输入门和候选信息的权重矩阵，$\tilde{C}_t$ 和 $b_c$ 则表示输入门和候选信息的偏置。

在通过输入门获取了新的信息后，根据公式(4)来更新当前时刻的记忆信息 $C_t$。

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \tag{4}$$

其中，$C_{t-1}$为上一时刻的记忆单元保存并传递过来的信息。

信息更新后便是信息的输出，主要由输出门来完成。输出门分为两个部分，首先根据输入信息 $x_t$ 和 $h_{t-1}$ 来获得门函数 $o_t$，如公式(5)所示：

$$o_t = \sigma\left(W_o \cdot \left[x_t, h_{t-1}\right] + b_o\right) \tag{5}$$

其中，$W_o$ 表示输出门的权重矩阵，$b_o$ 表示输出门的偏置。

其次，将上一步获取的 $C_t$ 与门函数 $o_t$ 相乘，从而得到本单元的的输出 $h_t$，如公式(6)所示：：

$$ht = ot \times \tanh\left(Ct\right) \tag{6}$$

正是由于以上的三种门结构使得 LSTM 网络能够很好的控制信息在记忆单元之间的流通，在许多自然语言处理问题中表现优秀[12]。

基于 LSTM 的文本情感分类方法流程图如图 2 所示。首先，将从社交平台上爬取的评论数据按 7:3 比例分为训练集与测试集。其次，将经过文本预处理和文本向量化操作后的文本数据导入神经网络，形成 Embedding 层。之后再进入神经网络的 LSTM 层，在这一层可以对窗口大小、迭代次数、等超参数进行调节，来优化模型的训练，同时为了防止模型产生过拟合，该层也添了 Dropout 方法，从而避免导致过拟合。

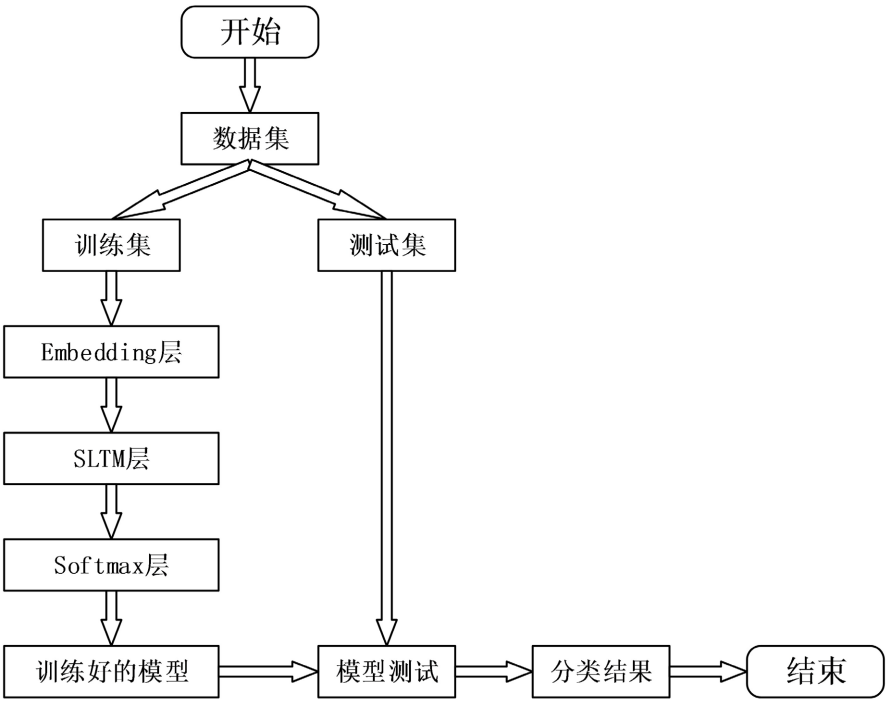


**Figure 2.** Flow chart of text sentiment classification method based on LSTM
**图 2.** 基于 LSTM 的文本情感分类方法流程图

## 4. 实验结果与分析

### 4.1. 实验数据

实验所用的设备与环境如表 1 所示。

**Table 1.** Experimental environment
**表 1.** 实验环境

| 工具 | 参数 |
|---|---|
| 操作系统 | Windows10 |
| CPU | Intel(R) Core (TM) i7-7700HQ |
| 显卡 | NVIDIA GTX 1050ti |
| 内存 | DDR4 16GB |
| 深度学习工具包 | TensorFlow/Keras |

本文的语料数据集来自亚马逊的审查数据集，其主要信息如表 2 所示。该数据集共有 10,000 条样本，其中积极情感样本 5097 条，消极情感样本 4903 条，其中 70%的文本样本作为训练数据集，30%的文本样本作为测试数据集。

**Table 2.** Dataset
**表 2.** 数据集

| 数据集 | 正类样本 | 负类样本 |
|---|---|---|
| 训练数据集 | 3568 | 3432 |
| 测试数据集 | 1529 | 1471 |
| 总数 | 5097 | 4903 |

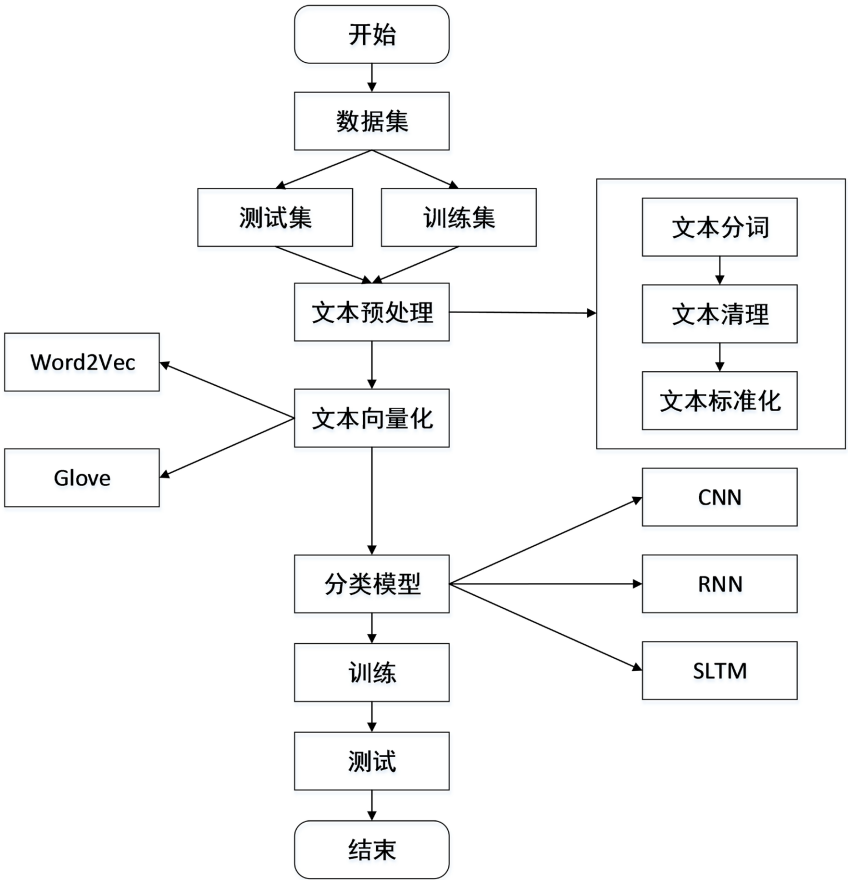


**Figure 3.** Flow chart of text sentiment analysis
**图 3.** 文本情感分析流程图

## 4.2. 实验流程与参数设置

文本情感分析实验分为四个步骤，如图 3 所示。

首先，从网络上整理数据获取数据集，并对数据集进行数据预处理，去除文本中存在的大量噪声，通常会先消除文本的格式，再对文本进行分词处理并删除无意义的停用词。将清洗好的数据采用 Word2Vec 算法或 Glove 算法进行文本特征提取，获取低维稠密的词向量。然后，建立文本情感分析模型，将词向量或句向量输入模型中进行训练。最后，将测试集输入到模型中获取测试结果。

本文采用 CNN、RNN 和 LSTM 模型分别进行了实验，三种模型所使用的参数如表 3 所示。

**Table 3.** Parameter setting
**表 3.** 参数设置

| 模型 | 参数 | 值 |
|---|---|---|
| CNN | 卷积层卷积核大小 | 5 |
| | 卷积核个数 | 128 |
| | 卷积层激活函数 | relu 函数 |
| | 全连接层单元个数 | 10 |
| | 全连接层激活函数 | relu |
| | 输出层损失函数 | binary_crossentropy |
| | 舍弃神经元占总神经元的比例 dropout | 0.5 |
| | 迭带次数 epoch | 10 |
| | 一次训练所选取的样本数 batch_size | 10 |
| RNN | 舍弃神经元占总神经元的比例 dropout = 0.5 | |
| | 激活函数 activation | tanh 函数 |
| | 迭代次数 epoch | 10 |
| | 一次训练所选取的样本数 batch_size | 64 |
| LSTM | 舍弃神经元占总神经元的比例 dropout | 0.5 |
| | 激活函数 activation | tanh 函数 |
| | 迭代次数 epoch | 4 |
| | 输出维度 units | 3 |
| | 一次训练所选取的样本数 batch_size | 32 |
| | LSTM 层的单元个数 output_dim | 50 |

## 4.3. 实验结果与分析

**Table 4.** Classification results
**表 4.** 分类结果

| 方法 | 准确率 | 精确率 | 召回率 |
|---|---|---|---|
| 无预训练方法 + CNN | 73.81 | 72.63 | 72.59 |
| 无预训练方法 + RNN | 83.76 | 82.92 | 82.05 |
| 无预训练方法 + LSTM | 85.30 | 84.91 | 84.47 |
| Word2Vec + CNN | 79.26 | 78.92 | 78.31 |
| Word2Vec + RNN | 89. 42 | 88.74 | 87.21 |
| Word2Vec + LSTM | **93.86** | **92.53** | **93.04** |

对本文数据进行情感二分类的实验结果如表 4 所示，观察表中的数据可以发现，基于 LSTM 的情感分类方法各项指标表现最优，说明在文本情感分析问题中，LSTM 模型本身的特点保证了情感分类的准确性和实时性。通过对比表 4 中各个模型的结果可以看出，基于 CNN 的方法效果并不理想，主要原因在于文本数据是存在密切联系的序列结构，卷积操作所产生的负面影响可能会高于其提取特征的效果，从而导致整体效果较差。而 RNN 和 LSTM 是专门处理序列数据的网络结构，保留了词与词之间的位置信息，因此 RNN 和 LSTM 的分类效果优于基于 CNN 的方法。然而，RNN 模型由于只能获取较短时间的信息，而基于 LSTM 的情感分类方法在 RNN 的基础上增加了记忆单元，能够有效解决长距离依赖问题，从而记住长期的信息。因此，基于 LSTM 的情感分类方法性能更好，各项评价指标均为最优。

## 5. 结论

针对 RNN 在处理长距离依赖问题时的不足，本文提出了基于 LSTM 的情感分类方法，在隐藏层添加记忆单元以达到长久记忆信息的目的。采用三种模型对文本情感倾向进行分类。实验结果表明卷积神经网络容易丢失文本的结构信息导致各项指标偏低。而 RNN 和 LSTM 保留了词与词之间的位置信息，在处理序列信息时具有较大优势。在三种模型中，基于 LSTM 的情感分类方法由于添加记忆单元，可以很好地解决长距离依赖问题，具有较好的分类效果。

今后可以进一步研究多分类模型，对文本情感进行更细致的分析。同时，研究更有效的分类方法，从而提高模型对于情感倾向的识别精度。

## 基金项目

大学生创新训练项目(S202110500067)。

## 参考文献

[1] Studiawan, H., Sohel, F. and Payne, C. (2020) Anomaly Detection in Operating System Logs with Deep Learning-based Sentiment Analysis. *IEEE Transactions on Dependable and Secure Computing*, **18**, 2136-2148. https://doi.org/10.1109/TDSC.2020.3037903

[2] 张军阳, 王慧丽, 郭阳, 等. 深度学习相关研究综述[J]. 计算机应用研究, 2018, 35(7): 1921-1928+1936.

[3] 王天笑. 自然语言处理的现状研究与未来发展初探[J]. 中国科技纵横, 2017, 1(2): 196-197.

[4] 刘爽, 赵景秀, 杨红亚, 等. 文本情感分析综述[J]. 软件导刊, 2018, 17(6): 1-4+21.

[5] Zhai, G., Yang, Y., Wang, H., *et al*. (2020) Multi-Attention Fusion Modeling for Sentiment Analysis of Educational Big Data. *Big Data Mining and Analytics*, **3**, 311-319. https://doi.org/10.26599/BDMA.2020.9020024

[6] 王根生, 黄学坚. 基于Word2vec和改进型TF-IDF的卷积神经网络文本分类模型[J]. 小型微型计算机系统, 2019, 40(5): 1120-1126.

[7] Singh, L.G. and Singh, S.R. (2020) Empirical Study of Sentiment Analysis Tools and Techniques on Societal Topics. *Journal of Intelligent Information Systems*, **56**, 379-407. https://doi.org/10.1007/s10844-020-00616-7

[8] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing*, Doha, October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[9] Sun, J., Han, P., Cheng, Z., *et al*. (2020) Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis. *IEEE Access*, **8**, 211152-211163. https://doi.org/10.1109/ACCESS.2020.3039470

[10] 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.

[11] 杨小平, 张中夏, 王良, 等. 基于 Word2Vec 的情感词典自动构建与优化[J]. 计算机科学, 2017, 44(1): 42-47.

[12] 何炎祥, 孙松涛, 牛菲菲, 等. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报, 2017, 40(4): 773-790.