基金项目论文

基于 SVM 和 LSTM 两种模型的商品评论情感分析研究

彭丹蕾，谷利泽，孙 斌

（北京邮电大学 网络空间安全学院，北京 100876）

摘 要： 随着网购的盛行，商品评论数量急剧增长，内容也越来越五花八门。如何高效挖掘处理这些评论是一件非常有价值的事情。对商品评论做情感分析是关于这些评论研究的一个重要方向。现阶段在情感分析研究中最常用的有基于机器学习的方法和基于情感知识分析的方法。本文主要采用机器学习中的 SVM 方法和深度学习中的 LSTM 方法分别对从京东网站爬取的商品评论进行模型搭建，然后对比分析。由于 LSTM 能够保持长期的记忆性，它很好地克服在 SVM 分类中每个句子的词向量求平均丢失了句子词语之间的顺序信息的缺点，保留了词与词之间的语义信息（如词序信息、上下文信息等），并且通过复杂的非线性计算更好地提取词向量中隐藏的情感信息。因此使用 LSTM 方法准确率比 SVM 方法提高不少，在情感分析上表现出非常好的效果。

关键词： 商品评论；情感分析；SVM；LSTM

中图分类号： TP181 **文献标识码：** A **DOI：** 10.3969/j.issn.1003-6970.2019.01.009

本文著录格式： 彭丹蕾，谷利泽，孙斌. 基于 SVM 和 LSTM 两种模型的商品评论情感分析研究[J]. 软件，2019，40（1）：41–45

Sentiment Analysis of Chinese Product Reviews Based on Models of SVM and LSTM

PENG Dan-lei, GU Li-ze, SUN Bin

(*School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing*, 100876)

【Abstract】: With the popularity of online shopping, the number of product reviews has increased dramatically, and its contents are becoming more and more diverse. How to efficiently mine these reviews is a very valuable thing. Emotional analysis of product reviews is an important aspect of these reviews. The most commonly used methods in sentiment analysis are machine-based learning and sentiment knowledge analysis at present. In this paper, SVM method in machine learning and LSTM method in depth learning are used to model the product reviews crawled from Jingdong website. Because LSTM method can maintain long-term memory, it can overcome the shortcoming of losing the order information between words in each sentence by SVM method, so the accuracy of LSTM method in test set is much higher than that of SVM method.

【Key words】: Product reviews; Sentiment analysis; SVM; LSTM

0 引言

随着电子商务的快速发展，网购已经被越来越多的人接受。它在给人们带来方便体验、低价产品的同时，也受到了由于无地域限制导致的购物质量匮乏，远距离鉴别困难，商品在网站的描述信息与现实实物不符等多种问题。所以，人们在网购某商品时越来越依赖已购客户对此商品的评价。但由于商品评论数量急剧增长以及内容五花八门，使得人们很难迅速准确地获取有价值的信息。所以如何有效挖掘处理这些评论显得非常重要。关于这方面的研究有很多，其中一个重要方向就是对这些商品评论做情感分析[1]。

情感分析又称评论挖掘或意见挖掘，是指通过

基金项目： 国家科技重大专项(批准号：2017YFB0803001)

作者简介： 谷利泽(1965–)，男，教授，主要研究方向：密码学，态势感知分析，网络舆情分析；孙斌，女，副教授，主要研究方向：云计算，网络舆情分析；彭丹蕾(1992–)，女，研究生，主要研究方向：网络舆情分析。

自动分析某种商品评论的文本内容，发现人们对这种商品的好评差评。现阶段在情感分析领域常用的有基于情感知识的方法和基于机器学习的方法[2-4]。基于情感知识的方法主要是利用一些已有的情感词典和语言知识对评论的情感倾向进行分类，包括 SentiWordNet、General Inquire、POS tragger 等等。它主要是以自然语言处理为基础，但因为现在 NLP 领域还存在着很多尚未攻克的难关。本文将研究基于机器学习的情感分析方法，具体选择的是机器学习中的支持向量机（Support Vector Machine, SVM）[5]。随着深度学习目前已经在图像处理和语音识别等领域得到了广泛地应用，它开始运用到情感分析研究中。因此本文也将研究基于深度学习的情感分析方法，具体选择的是深度学习中的长短时记忆网络（Long Short-Term Memory，LSTM）[6]。通过 SVM 和 LSTM 结果对比，得出结论。

1 相关知识介绍

1.1 支持向量机（SVM）

支持向量机[7]（SVM）是一种有监督学习方法，以结构风险最小化为原则，是一种具有很好泛化能力的分类工具，目前已经被广泛应用于文本分类以及人脸识别等多个领域。它是一种二分类模型，基本模型就是定义在特征空间上的间隔最大的线性分类器。它主要包含线性可分线性可分支持向量机模型、线性支持向量机模型以及非线性支持向量机模型。线性可分支持向量机模型指训练数据线性可分时，通过寻求硬间隔最大化构建的线性分类器；线性支持向量机模型指训练数据接近线性可分时，通过寻求软间隔最大化构建的线性分类器；非线性支持向量机模型指训练数据线性不可分时，通过运用核技巧和软间隔最大化构建的非线性分类器模型（其中核技巧是指巧妙地利用线性分类学习方法与核函数解决非线性问题的技术）。

关于 SVM 的基本模型[8]。设训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中，$x_i\in R^n$，$y_i\in\{-1,+1\}$，i=1,2,⋯,N，满足：

$$y[i](w\cdot x[i]+b)\geq 1 \quad (1)$$

并使得

$$\min\|w\|\times\|w\|/2 \quad (2)$$

根据拉格朗日对偶性，通过线性可分支持向量机的对偶算法即求解对偶问题来得到原始问题的最优解。转换后为：

$$\max\sum\alpha[i]-1/2\sum\alpha[i]*\alpha[j]*y[i]*y[j]*x[i]*x[j] \quad (3)$$

$$\text{s.t. } 0\leq\alpha[i]\leq C, a[i]*y[i]=0,\ i=1,2,3,\cdots,N \quad (4)$$

上式中 $x[i]*x[j]$ 代表这 2 个向量的内积，当遇到线性不可分的时候，用核函数 $K(x[i]*x[j])$ 代替 $x[i]*x[j]$。通过对偶问题的解 α，求得 w 和 b，从而得到最大间隔分离超平面以及分类决策函数。

1.2 Word2Vec

Word2vec 是 Google 在 2013 年推出的一款将所有词向量化的 Deep Learning 工具，是面向大众的开源的。传统的词向量表示方法主要是 One-hot Representation。它的思想是词表中的词都可以用一个长向量进行表示，向量的维度就是词表的大小，只需将词在词表中的位置所对应长向量的位置置为 1，其余的都为零。虽然简单，但很容易造成“维度灾难”，还有词语之间的相似性无法反映。基于神经网络的语言模型就应运而生。Word2vec 就是其中最常见的一种，思路主要是通过神经网络训练实现将文本中的词转换为词向量，然后映射到新的空间上去[9]。词语之间的相似度主要通过计算映射到新空间上向量之间的余弦相似度来衡量。它不需要标注数据，自发学习词词之间的某些内在联系目前因此已被广泛的运用到自然语言处理领域中。

Word2vec 一般分为 CBOW 和 Skip-gram 两种模型。CBOW 模型主要是通过某一个特征词周围的词所对应的词向量预测这个特征词的词向量：

$$P(W_t\,|\,W_{t-k},W_{t-k+1},\cdots,W_{t+k-1},W_{t+k}) \quad (5)$$

W_t 代表当前语料中的一个特征词。通过和 W_t 相近的 k 个词来预测词 W_t 出现概率情况。Skip-gram 模型与 CBOW 恰好相反，它是通过一个特征词的词向量，来预测周围词所对应的词向量：

$$P(W_{t-k},W_{t-k+1},\cdots,W_{t+k-1},W_{t+k}\,|\,W_t) \quad (6)$$

Word2vec 通过 CBOW 和 Skip-gram 两种模型来训练得到相应词向量。输入文本来预测每个单词对应的词向量。

1.3 长短期记忆网络(LSTM)

LSTM（长短期记忆网络）[10]是 RNN（循环神经网络）的一种特殊的类型，主要克服了 RNN 无法解决长期依赖的缺陷。RNN 的优点在于不仅会学习当前时刻的信息，而且还会依赖之前的信息。RNN 特殊的网络模型结构也有效地解决了信息保存的问题。所以 RNN 在处理时间序列和语言文本序列问题上有着广泛的应用。但是因为多数的 RNN 结构训练

时候都存在着梯度消失，梯度爆炸等问题。为了避免 RNN 这些缺陷，能够很好地控制在训练过程中梯度的收敛性，同时也可以保持长期记忆性提出了 LSTM。LSTM 通过细胞的记忆单元来替代 RNN 隐藏层中的模块，同时使用输入门、输出门、遗忘门使得其可以记忆、更新长距离的信息，从而实现对长距离信息的处理。LSTM 的具体架构图如下。

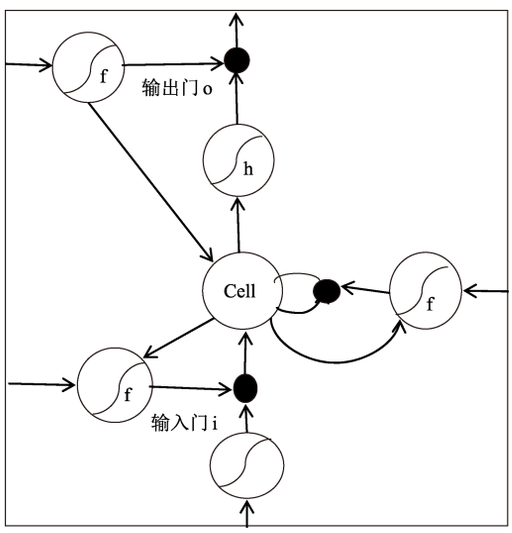

图 1　LSTM 结构图
Fig.1　LSTM architecture diagram

LSTM 主要由输入门 i、输出门 o、遗忘门 f 和记忆细胞 c 组成。输入层、输出层和遗忘层分别控制记忆细胞的读、写和丢失操作的控制器。

图中 Cell 是整个 LSTM 的关键，它类似于传送带，直接在整个循环网络中运行，并且由它来决定该保留或者遗弃哪些信息。而其他的三个层，则用来保护或者遗弃哪些信息。而其他的三个层，则用来保护和控制细胞的状态，最后由输出层的输出 o_t 以及当前时刻 cell 的状态 c_t 决定整个 LSTM 的输出 y_t。LSTM 可以表述为：

$$z^t = \tanh(W_z x^t + R_z y^{t-1} + b_z) \quad (7)$$

$$c^t = i^t z^t + f^t c^{t-1} \quad (8)$$

$$y^t = o^t h(c^t) \quad (9)$$

$x^t \in R^d$ 为输入矢量，$y^t \in R^d$ 为 t 时刻 LSTM 的含有语句特征的输出矢量。δ，h 表示激活函数 sigmoid 和 tanh，W_*, R_*, p_*和b_*分别代表系数矩阵和偏置向量。

1.4　TensorFlow

TensorFlow[11]分成两部分张量（tensor）和流（flow），张量是计算图的基本数据结构，它代表的是节点之间传递的数据，通常是一个多维度矩阵或者一维向量，流表达了张量之间通过计算互相转化的过程，形象理解为数据按照流的形式进入数据运算图的各个节点。目前，tensorflow 是最热门的深度学习工具，本文将借助 tensorflow 实现深度学习 LSTM 模型。

2　具体实现

2.1　实验数据集

本文所用的主要实验数据集为京东网站爬取的商品评论文本以及一些公开数据集。其中爬虫主要是借助 Scrapy 框架。Scrapy 是一个为了爬取网站数据，提取结构性数据而编写的应用框架。可以应用在包括数据挖掘，信息处理，监测和自动化测试等领域。本文使用 Scrapy 获取商品评论文本过程如下：

（1）在终端命令行 cmd 输入 scrapy startproject JDview，就创建了一个 scrapy 的爬虫工程。

（2）在 items.py 定义抽取的字段，本次只有 content 这一项内容。

（3）编写 spider，定义一个用于下载的初始 url，即 start_urls，对页面的内容进行解析，提取在 items.py 规定需要的字段，同时将结果保存到 csv 文件中。

（4）由于在第 3 步实现了文件存取，故没有修改 pipelines.py。

数据集的选取

由于京东商城是国内最大的数码通讯、电脑和家用电器的网上购物商城，同时，该网站也是中国电子商务领域中最具有影响力的电子商务网站之一。故本文爬取部分五星级好评和一星级差评，再加上网上的一些公开商品评论中文标注语料共同构成数据集。这些数据集共有两万多条，涵盖多个领域，手机，电脑，书籍，食品等。其中正负比例大约 1∶1。通过整合整理将所有正的保存在 pos.txt，将所有负的保存在 neg.txt。

2.2　使用 SVM 进行情感分类

主体思路：从数据集中抽取样本，构建比例为 8:2 的训练集和数据集。随后，对训练集数据构建 Word2Vec 模型。其中分类器的输入值是一条条评论所有词向量的加权平均值。Word2Vec 工具和 svm 分类器分别使用 python 中的 gensim 库和 sklearn 库[12]。

（1）将正负两组数据利用 python 的 numpy 进行整合，同时按照它们的顺序生成相应的标签，正（积极情绪）用 1 表示，负（消极情绪）用 0 表示。同

时对正负每条评论分别利用 jieba 分词工具分词，进行下简单的词性筛选，利用 sklearn 包下的 train_test_split 函数，构建比例为 8:2 的训练集和测试集。

（2）利用 gensim 库一些关于 word2vec 函数来实现计算词向量，例如用 gensim.Word2Vec 来训练词向量模型，也需要定义一些参数，size 就是一个很重要的参数，指的是输出词向量的维数。定义好的模型通过 build_vocab 以及 train 与之前的训练集关联，来实现对训练集数据构建 Word2Vec 模型。之后我们对每个句子的所有词向量取均值来作为分类器的输入值。

（3）实现训练 SVM 模型。具体用的是 sklearn 包下的 SVC 函数，它是基于 libsvm 来实现的。其中 kernel 参数指定用的是 rbf。之后调用其下的 fit 函数来具体训练 SVM 模型。最终调用 score 函数用测试数据集来对已建好的 SVM 模型进行检验来判定模型的准确率。

2.3　使用 LSTM 进行情感分类

主题思路：分为两部分，一是输入数据处理，由于神经网络无法处理字符串，因此需要将字符串转换为数字，并进行一定处理将其输入后续模型。二是 LSTM 模型构建，使用 python 语言借助 tensorflow 框架来实现模型搭建，比如 tensorflow(tf) 的 tf.contrib.rnn 下的 BasicLSTMCell，MultiRNNCell，dynamic_rnn 等函数[13]。

（1）将文本转换为数字，具体方法就是给每个单词贴上一个数字下标，同时将数据集中的每一条评论从字符串转为数字。同时按照数据集的评论根据正负生成相应的标签，正（积极情绪）用 1 表示，负（消极情绪）用 0 表示。

（2）建立 LSTM 模型，首先设置诸如 LSTM 个数，层数等基本参数，通过 tensorflow 的 placeholder 来定义输入输出。通过 tf.nn.embedding_lookup 添加 Embeding 层，来实现之前类似 word2vec 模型的功能。接着就是具体的建立 LSTM 层，借助的是 tf.contrib.rnn 下的 BasicLSTMCell，MultiRNNCell，dynamic_rnn 等函数。然后定义输出，验证准确率，这些就是模型的基本模块。

（3）具体训练，主要是将数据集和模型进行连接，以 batch 为单位输入神经网络，来进行训练。同时在基本参数将训练次数设置为 10，在最终输出是每 5 个 batch 输出一次损失率，每 25 个 batch 输出一次准确率。

3　实验结果展示

3.1　SVM 模型结果

图 2 是 SVM 运行的结果。

从图中可以直观的看出，针对那些商品评论训练出的 SVM 模型的准确率为大约 0.814，效果整体还是不错。通过多次运行程序，程序输出的准确率基本都稳定在这个值左右。

3.2　LSTM 模型结果

图 3 是 LSTM 运行的结果。

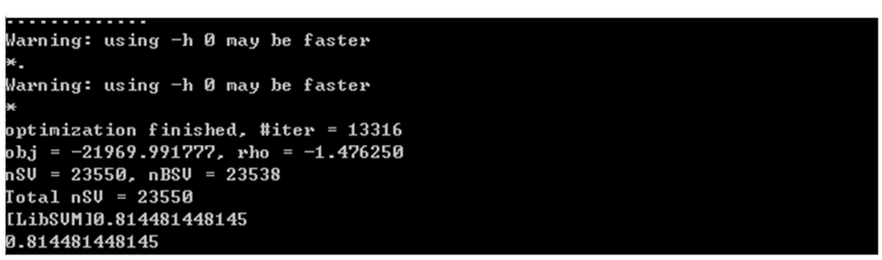

图 2　SVM 运行结果图

Fig.2　SVM Operation result diagram

从上图可以看出在 LSTM 模型训练第一轮 (Epoch:0/10)的时候损失率由大约 0.248 到后来不断优化，变成约为 0.206。到第十轮(Epoch /10)后损失率已经变得很小，基本是在 0.01 至 0.03 之间。准确率也是由第一轮最开始的 0.54，一直优化到接近最后的 0.900。准确率也不断提高。经过多次运次程序，每次程序经过 10 轮迭代后准确率输出也基本都在 0.900 左右。

3.3　SVM 与 LSTM 模型对比

通过截图就可以直观地看出 LSTM 模型的准确率比 SVM 模型的准确率高一些，同时多次运行结果稳定。根据在相关知识介绍可知 SVM 虽然在二分类问题上有一定的优势，但由于评论语句每个词之间并不是一点关系也没有，而本文在构建情感模型的时候，把每个句子的词向量求平均作为输入丢失了句子词语之间的顺序信息。LSTM 在构建文本分析这类需要长期依赖学习模型时，优于其他模型，而且由于 LSTM 模型遗忘、记忆和更新信息的能力使得它领先 RNN 一步(LSTM 模型是特殊的一种 RNN 模型)。

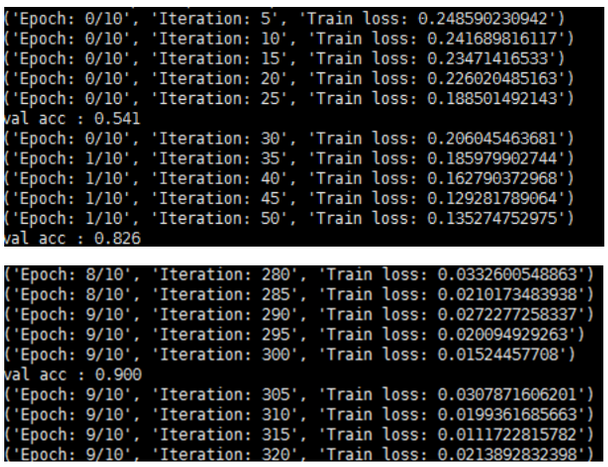

图 3　LSTM 运行结果
Fig.3　LSTM Operation result diagram

4　总结与展望

随着互联网进入千家万户，人们在网上的活跃度也越来越高，对新闻进行讨论，对产品进行评价成为了人们互联网生活的一部分。互联网上的语言越来越多样、随意，迫切需要新的技术来弥补传统情感分析方法的局限。本文提出了基于 SVM 和 LSTM 两种模型来对中文商品评论进行情感分析。主要贡献有三点：

（1）使用机器学习中常用的 SVM 方法来实现情感分析二分类问题。主要抽取样本构建一定比例的训练集和测试集。随后，对训练集数据构建 Word2Vec 模型，其中分类器的输入值为每一条评论中所有词向量的加权平均值。Word2vec 工具和 svm 分类器分别使用 python 中的 gensim 库和 sklearn 库。

（2）使用深度学习中常用的 LSTM 方法来实现情感分析。主要是借助 tensorflow 来搭建 LSTM 模型，同时词语向量化主要是借助 tensorflow 下的 embedding_lookup。

（3）对比针对同样的数据集，同样的数据预处理方法，两个模型准确率的差异，得出 LSTM 模型在学习长距离依赖，以及捕捉文本的序列特征与上下文依赖关系有很大的优势。

情感分析涉及到的学科多，而且学科交叉的跨度大，本文只是从一小点进行探索。在进一步研究中，一方面，需要在更广泛的数据集上验证本研究的结论，例如引入英文评论，另一方面，也需要对 LSTM 做更深入的研究，包括现在已经出现了很多 LSTM 的优化模型。

参考文献

[1] 钟将，杨思源，孙启干．基于文本分类的商品评价情感分析[J]．计算机应用, 2014, 34(8): 2317-2321.

[2] 张紫琼，叶强，李一军．互联网商品评论情感分析研究综述[J]．管理科学学报, 2010, 13(6): 84-96.

[3] Scholz T, Conrad S. Linguistic Sentiment Features for Newspaper Opinion Mining[M]//Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2013: 272-277.

[4] 赵妍妍，秦兵，刘挺．文本情感分析[J]．软件学报，2010, 21(8): 1834-1848.

[5] 樊康新．基于SVM的网络文本情感分类系统的研究与设计[J]．计算机时代, 2015(12): 34-37.

[6] Jun, Liang, et al. "Polarity Shifting and LSTM Based Recursive Networks for Sentiment Analysis." *Journal of Chinese Information Processing*(2015).

[7] Harrington P. Machine Learning in Action[M]. Manning Publications Co, 2012.

[8] 樊康新．基于SVM 的网络文本情感分类系统的研究与设计[J]．计算机时代, 2015(12): 34-37.

[9] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[10] 陈葛恒．基于极性转移和双向LSTM 的文本情感分析[J]．信息技术, 2018(2):149-152.

[11] Abadi M. TensorFlow: learning functions at scale[J]. Acm Sigplan Notices, 2016, 51(9): 1-1.

[12] Zhang D, Xu H, Su Z, et al. Chinese comments sentiment classification based on word2vec and SVM perf[J]. Expert Systems with Applications, 2015, 42(4): 1857-1863.

[13] Tang D, Qin B, Feng X, et al. Effective LSTMs for Target-Dependent Sentiment Classification[J]. Computer Science, 2015.