

毕业设计说明书

|  |  |  |  |
| --- | --- | --- | --- |
| **作 者:** | 赵斯蒙 | **学 号：** | 913106840231 |
| **学 院:** | 计算机科学与工程学院 | | |
| **专业(方向):** | 计算机科学与技术 | | |
| **题 目:** | 基于卷积神经网络的文本的情感分析 | | |
|  |  | | |

|  |  |
| --- | --- |
| **指导者：** | 贾修一 副教授 |
|  |  |
| **评阅者：** | \*\*\* 副教授 |

2017 年 5 月

声 明

我声明，本毕业设计说明书及其研究工作和所取得的成果是本人在导师的指导下独立完成的。研究过程中利用的所有资料均已在参考文献中列出，其他人员或机构对本毕业设计工作做出的贡献也已在致谢部分说明。

本毕业设计说明书不涉及任何秘密，南京理工大学有权保存其电子和纸质文档，可以借阅或网上公布其部分或全部内容，可以向有关部门或机构送交并授权保存、借阅或网上公布其部分或全部内容。

学生签名：

年 月 日

指导教师签名：

年 月 日

# 绪论

卷积神经网络作为近几年再次兴起的机器学习算法，受到了学术界的广泛关注，随之而来的深度学习热潮以及其在各个领域上的应用，更是为人类掀起了了技术革命的新篇章

## 工程背景及问题

短文本的分类一直是机器学习领域一个比较困难的问题，由于短文本自身所具有的特点：上下文信息匮乏、字词的歧义性，致使短文本的分类本身就具有一定的不确定性，而微博文本，微博评论就是典型的短文本，对此类短文本进行情感分析以及更进一步的情感分类，对于舆情研究，满意度反馈，证券投资[[1]](#endnote-1)等领域，具有十分巨大的潜力。对于此类文本，传统文本分析的方法无法提供足够的精确度，面对微博文本这种时效性和口语性较强的文本类型时，更是无法给出让人足够满意的结果，所以，利用机器学习方法对此类短文本并进行分类就成为了一个潜力巨大的新方向。

## 相关技术的现状

目前，实现文本分类的机器学习技术主要有支持向量机(SVM)[[2]](#endnote-2)，朴素贝叶斯[[3]](#endnote-3)，以及本文中提到的神经网络等，在实际使用中，也有结合了传统的词典分类以及情感标签[[4]](#endnote-4)的分类方法，都取得了一定的成效，随着机器学习领域的不断进步以及近几年卷积神经网络和深度学习的发展，对于短文本分析的精度被不断刷新；除了分类器(classifier)的算法为文本分类提供了有力支撑，随着词向量[[5]](#endnote-5)的出现，文本数据有了更好的特征选择和特征表示方法；中文文本分词技术的提高则更好的利用了词向量所带来的语义上的优势；此外，计算能力的提升数据源的增加也为卷积神经网络在文本分类上的应用的成功做出的不可忽视的贡献。

## 需解决的工程问题

要实现文本分类，首先想到的就是如何提取特征，使要输入的文本变成可以被模型所识别的数值形式，鉴于近年来词向量在文本分类领域的优异表现，本实验中决定采用的特征提取方式就是词向量，但是从源文本转化成词向量还需经历两个步骤，分别是“词”和“向量”，显然，中文文本和英文文本有着明显的区别英文以及其他诸多语言都有着天然的分词模式，而中文的分词则没有这种特点，有些时候甚至会出现一词多义、一句多分的情况，这就需要在特征选择阶段对文本进行分词，随着时代的变化，更多新名词的出现，分词的工具也需要不断更新迭代；在分词结束后，要给每一个分好的词映射对应的向量值，根据词向量5的解释，在获得实际向量值之前还需要另外训练一个词向量的模型，从而得到每一个词对应的向量值；在得到了一个句子的每一个词的向量值后，就可以近似的把这个句子作为一个图片，至此我们就得到了一个句子的数值矩阵形式，而且这种形式能很好的反映出这个句子所具有的语义特征；得到了可以输入的数据之后，接下来的工作就是把数据置入模型进行训练，但是要进行训练现在的数据维度还是太大，为了解决这个问题，我们引入了卷积操作，在合适的卷集合的大小之下，目标文本的序列特征将得到足够的体现。但是在进入神经网络模型训练前，还需要进行一个步骤，因为神经网络对于输入数据的大小有确定性要求，所以需要进行池化(pooling)操作，使每一个卷积操作的结果只有一个会被传输到输入端。在上述步骤完成后，到达神经网络的输入端的数据一定是确定数量的卷积核的输出值，这个确定的数量就是卷积核的数量；之后，要搭建适合的神经网络，在搭建神经网络的过程中，要选择适当的训练算法和学习率等超参数，以较好的适应学习的情景。在训练的过程结束之后，就需要进行验证的步骤，对模型的准确率进行评估，所以，在训练之前还要注意原始数据集的划分工作，划分成用于训练的训练集和用于测试的测试集。

## 论文章节安排

本篇论文主要介绍了借助卷积神经网络对微博文本进行情感分析的理论基础和系统流程，

# 卷积神经网络原理

神经网络算法最早源于上世纪40年代，其主要思想是模仿人类神经元的工作特性，每一个输出值都与每一个输入值呈线性关系，通过增加模型的复杂度和参数的数量，增加模型的精确性和可靠性，随着计算能力的增加，可以训练的神经网络的规模也越来越大，在各个机器学习的应用领域取得了越来越好的结果。

## 神经网络的基本原理

……

f

人工神经网络(artificial neural network)ANN，简称神经网络，最基本的单位是神经元：

其中，神经元的每一个输入都乘以相应的权重并

## word embedding得到词向量

# 算法设计

卷积神经网络由两部分组成，分别是卷积层和神经网络层，其中，神经网络的设计源于人类神经元的突触，激活的方式，让各个神经元之间彼此相连，通过调整参数来实现模型的拟合，而卷积层的操作可近似作为一个参数稀疏的神经网络层，其主要功能是降维和挖掘相邻像素(在本例中是词汇)之间的关系。

## 算法的基本原理

### 词向量

词向量的概念最早出现于

### 卷积神经网络

### 梯度下降算法

## 推导公式和理论基础

### 神经网络

### word to vector词向量模型

### 卷积神经网络

### 交叉熵与权重衰减

### 梯度下降算法

## 算法实现

# 系统实现以及测试

## 框架搭建

采用python下的tensorflow框架，（简要介绍运行环境，tensorflow 的工作方式，）

## 数据获取

## 数据预处理

1. 陆宇杰. 中文微博情感分析及其应用[D]. 华东师范大学, 2013. [↑](#endnote-ref-1)
2. Mccallum A, Nigam K. A Comparison of Event Models for Naive Bayes Text Classification[J]. IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION, 1998, 62(2):41--48. [↑](#endnote-ref-2)
3. Tong S, Koller D. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research, 2002, 2(1):45-66. [↑](#endnote-ref-3)
4. 孙建旺, 吕学强, & 张雷瀚. (2014). 基于词典与机器学习的中文微博情感分析研究. *计算机应用与软件,* *31*(7), 177-181. [↑](#endnote-ref-4)
5. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model[J]. Journal of machine learning research, 2003, 3(Feb): 1137-1155. [↑](#endnote-ref-5)