

**内蒙古师范大学计算机科学技术学院**

**毕业设计（论文）开题报告**

**题 目：基于深度学习的文本分类**

**专 业 ：信息管理与信息系统**

**学 生 ：管贤玉**

**学 号 ：20161102967**

**指导教师 ：苗忠义**

**日 期 ：2019/11/21**

**计算机科学技术学院制**

**说 明**

1. 课题来源及研究的目的和意义

课题来源于想利用学过的知识和现在社会大数据分析的趋势，做一个python的爬虫来抓取一些文本数据做文本的情感分析。

现代文本的特征：现代社会已经进入大数据时代,数据挖掘和人工智能已成为时代主题。在各种领域和社会生活的方方面面,呈现出海量数据的特征。其中文档数据更是蕴含各种信息,如果对文本认真挖掘加以科学分析利用,将会产生极高的社会价值。

现代文本的获取：比如说一直在生活中扮演着重要角色的微博，微博下评论的人很多，但是也分精选评论，和所谓的恶评。什么是恶评，就需要一个模型来判断。还有一些发新闻的媒体通过用户对某一新闻评论的感情，来预备下一次新闻稿的情感走向。

文本的分类：文本分类可减少信息抽取的选择空间,并可针对不同的信息类型,制定不同的策略,是信息抽取不可或缺的前提步骤。希望通过文本的分类，做一些情感分析，判断文本是积极还是消极还是中立。

1. 文本分类和情感分析的研究现状及分析

中国计算机学会（CCF）主办，雷锋网，香港中文大学（深圳）承办的2018全球人工智能与机器人峰会（CCF-GAIR）在深圳召开， 国内做情感分析比较好的有：哈工大秦兵老师，秦老师多年从事自然语言处理的研究，获得国内第一个关于文本情感分析方面的自然科学基金重点项目大连理工林鸿飞老师，在细粒度情感分析方面做出了很多贡献，其中独立整理标注完成的情感词汇本体库。

现阶段对情感分析的数据库已经初步成型，但是还需要完善，因为人类的感情很复杂，我们经常能想起来的词或者看到的词，比如喜极而泣，抱头痛哭，捶胸顿足，七情六欲，五味杂陈等等，表达了我们的喜怒哀乐。实际上多年来也有很多人在这方面做了很多研究，比如七情六欲分为好，恶，乐，怒，哀，惧，欲等。此外，还有人从高兴，悲伤，愤怒，恐惧，厌恶，惊奇等进行分类。所以随着深度学习的发展，也给文本情感分类带来很多生机。这些年出现很多关于深度学习和情感分析相结合的情感分析。在深度学习过程当中可以加入情感资源包，加入之后可以使情感分类和深度学习结合得更充分。所以我希望做一些这个方面的完善。

1. 主要研究内容

文本分类问题： 选定一段文本，将文档分类为多个类别中的一个或多个；文本分类应用： 常见的有新闻关键字锡类，情感分析；文本分类方向： 主要有二分类，多分类，多标签分类；文本分类方法： 传统机器学习方法（贝叶斯，svm等），深度学习方法（fastText，TextCNN等）。

研究的内容致力让读者或者媒体明白某一段文本重点关注什么问题，对于不同的场景应该采用什么方法。

文本分类的处理大致分为文本预处理、文本特征提取、分类模型构建等。和英文文本处理分类相比，中文文本的预处理是关键技术。（分类来源于网络）

4. 研究方案

主要使用textcnn算法，

1.TextCNN 是利用卷积神经网络对文本进行分类的算法，由 Yoon Kim 在 “Convolutional Neural Networks for Sentence Classification” 一文中提出是2014年的算法。该算法在一系列文本分类任务（如情感分析）中实现了良好的分类性能，并已成为新的文本分类架构的标准基准。

2.准备好需要的库和数据集。爬取淘宝评论，或者网易新闻，存到一个数据集中。

3. 数据和预处理。对数据进行分类，去掉文字中的符号，表情等。

4. 建立模型。

5．进度安排，预期达到的目标

1）前期首先学习python爬虫的相关知识，了解网络的安全法，了解卷积神经网络和循环神经网络。

2）前中期了解网页的基本结构div，img，href的用法和位置，熟练操作x-path

3）中期爬取数据，利用机器学习库TensorFlow，分析文本内容。

4）中后期回顾学习过的经济学、统计学的知识，利用textcnn算法，编写程序，对文本内容进行具体的情感分析。建立一个情感分析库。

5）后期归纳总结文本的关键词，作为日后的情感分析的基础。

6．课题已具备和所需的条件、经费

电脑安装有pycharm、x-path、和谷歌浏览器，可以做简单的数据获取和数据分析，有相应软件的教学课程，书籍等。

7．研究过程中可能遇到的困难和问题，解决的措施

1）在python抓取网站数据的过程中可能会碰到相关网站的封号情况，或者是网站相关路径输入错误造成的数据抓取错误。

解决方案，设计访问网站的username，伪装进站，在代码中加入时间函数，仿照人为浏览网页的情况。网页数据，利用浏览器的开发者选项配合x-path进行数据的精确获取。

2）中文乱码问题：使用pycharm进行python文本分词时，得到的每个字都是乱码，更改数据编码方式。

3）对TextCNN模型原理及理解不足，造成的分类错误，因为其主要思想是将不同长度的短文作为矩阵输入，使用多个不同size的filter去提取句子中的关键信息（类似于多窗口大小的ngram），并用于最终的分类。解决措施，理解模型和原理。

8．主要参考文献

[1] Kim, Yoon. "Convolutional Neural Networksfor Sentence Classification." Proceedings of the 2014 Conferenceon Empirical Methods in Natural Language Processing (EMNLP) 2014

[2] 韦灵，《基于机器学习的中文文本自动分类的实践研究》-《智库时代》,2019年.46期

[3] 董杰盛，《自然语言处理在新闻分类中的应用》-《科学咨询(教育科研)》,2019年.11期

[4] 陶志勇，《基于双向长短时记忆网络的改进注意力短文本分类方法》文献来源：数据分析与知识发现

[5] 梁柯，《基于朴素贝叶斯的文本情感分类及实现》-《智能计算机与应用》,2019年.05期

[6] 孙嘉琪，《基于神经网络模型的文本分类研究综述》-《天津理工大学学报》,2019年.05期

[7] 于游，《中文文本分类方法综述》-《网络与信息安全学报》,2019年.05期

[8] Ian James Bruce Young; Saturnino Luz; Nazir Lone，《A systematic review of Natural Language Processing for classification tasks in the field of incident reporting and adverse event analysis》- http://doi.cnki.net/doi/Resolution/Handler?doi=%2010.1016/j.ijmedinf.2019.103971

[9] 陈文实，《面向多标签文本分类的深度主题特征提取》-《模式识别与人工智能》,2019年.09期

[10] 涂文博，《针对文本分类的神经网络模型》-《计算机系统应用》,2019年.07期

[11] 李文江，《基于深度学习的商品评论情感分类研究》-《知识管理论坛》,2018年.06期

[12] 刘春磊，《基于TextCNN的用户评论情感极性判别》-《电子世界》,2019年.03期