北京理工大学

本科生毕业设计（论文）开题报告

**学 院：** 计算机学院

**专 业：** 计算机科技与技术

**班 级：**

**姓 名：** 苏尼

**指导教师：** 黄永刚

**校外指导教师：** 黄永刚

二○二一年二月八日

毕业设计（论文）开题报告评审表

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 姓名 | | 苏尼 | | 学号 | 1820171013 | 班级 | 07111703 | 专业 | | | 计算机科技与技术 |
| 导师 | | 黄永刚 | | 校外导师（职称） | | 黄永刚 副教授 | | 校外导师单位 | |  | |
| 论文选题 | 题目名称 | | | 基于深度学习的文本感情分析系统研究与实现 | | | | | | | |
| 题目性质 | | | 软件开发（ ✓ ） 理论研究（ ）  工程设计（ ） 技术科学研究与工程技术研究（ ） | | | | | | | |
| 题目来源 | | | 结合科研 （ ✓ ） 结合生产实际（ ）  结合实验室建设（ ） 自拟题目 （ ） | | | | | | | |
| 评审组成员 | 姓 名 | | | 职 称 | | 工作单位及职务 | | | 签 字 | | |
|  | | |  | | 北京理工大学 | | |  | | |
|  | | |  | | 北京理工大学 | | |  | | |
|  | | |  | | 北京理工大学 | | |  | | |
| 评  审  意  见 | （含：选题意义；选题是否满足毕业要求；技术方案是否可行；进度安排是否合理等） | | | | | | | | | | |
| 成 绩 | | |  | | | | | | | | |
| 评审组长签字： 年 月 日 | | | | | | | | | | | |

注：成绩以“合格”“不合格”记；评审组长为高级职称人员。

# 毕业设计（论文）选题的内容

在当今现代世界中，我们正遭受数据过载的困扰，公司经常收集有关客户反馈，购物行为等方面的大量数据。公司可以通过以下方式灵活地更改其数字资料，产品或服务，以最适合新市场和客户：分析这些数据。 但是，对于任何人来说，在没有任何错误或偏见的情况下，仍然很难手动解释它。情感分析是一种评估一段文字或文字是肯定的，否定的还是中立的方法。 情绪分析使跨国公司的数据分析人员能够评估公众情绪和产品感知，并考虑消费者的感知。

社交平台每天产生海量的文本数据，其中包含了丰富的价值信息，对其进行文本情感分析，有利于企业精准营销和消费者了解商品信息等。本课题主要基于目前主流的深度学习技术和Web技术，研究与实现一个基于深度学习的文本情感分析系统。

# 研究方案

* 1. 本选题的主要任务

了解文本情感分析领域背景知识，了解国内外技术发展趋势，理解其对环境以及社会可持续发展的影响，理解相关行业的政策和法律法规；

在指导教师指导下阅读国内外文献和自学相关知识，对基于深度学习的文本情感分析系统进行研究和实现。

主要包括：

1）深度学习算法模块：基于主流深度学习算法，训练文本情感分析模块；

2）服务器模块：接受前端请求，调用深度学习算法模块，并返回结果；

3）前端模块：用户交互；

* 1. 技术方案的分析、选择

1. 数据收集：

Twitter US Airline Sentiment数据集：是Twitter上美国航空公司情绪数据集 (Twitter US Airline Sentiment)，自 2015 年 2 月以来美国航空公司的 Twitter 数据，分类为正面、负面和中性推文。我们可以从github或者kaggle比赛官方网站进行下载。

1. 分析

我们本论文中将构建一个深度学习模型，将文本分类为否定或肯定。流程上，也会进行传统机器学习方法的探索和尝试，进行对比。

C． 流程

探索数据、预处理数据，对数据处理后，接下来开始训练模型、评估模型，然后优化模型等步骤。

下载好数据以后，一般不会急于创建模型、训练模型，在这之前，需要对数据、对需求或机器学习的目标进行分析，尤其对数据进行一些必要的探索，如了解数据的大致结构、数据量、各特征的统计信息、整个数据质量情况、数据的分布情况等。为了更好体现数据分布情况，数据可视化是一个不错方法。

这一块主要考虑可视化、统计各列和结果的相关性热力图等方法。

通过对数据探索后，可能发现不少问题：如存在缺失数据、数据不规范、数据分布不均衡、存在奇异数据、有很多非数值数据、存在很多无关或不重要的数据等等。这些问题的存在直接影响数据质量，为此，数据预处理工作应该就是接下来的重点工作，数据预处理是机器学习过程中必不可少的重要步骤，特别是在生产环境中的机器学习，数据往往是原始、为加工和处理过，数据预处理常常占据整个机器学习过程的大部分时间。

在模型选择时，一般不存在某种对任何情况都表现很好的算法（这种现象又称为没有免费的午餐）。因此在实际选择时，一般会选用几种不同方法来训练模型，然后比较它们的性能，从中选择最优的这个。当然，在比较不同模型之前，我们需要先确认衡量性能的指标，对分类问题常用的是准确率，对回归连续性目标值问题一般采用误差来评估。训练模型前，一般会把数据集分为训练集和测试集，或对训练集再细分为训练集和验证集，从而对模型的泛化能力进行评估。这一块我们会重点考察探索传统机器学习[K最近邻](http://telecom%20industry%20customer%20churn%20prediction%20with%20k%20nearest%20neighbor/)算法和深度学习算法LSTM来考虑时序的影响。

使用训练数据构建模型后，通常使用测试数据对模型进行测试，测试模型对新数据的测试。如果我们对模型的测试结果满意，就可以用此模型对以后的进行预测；如果我们测试结果不满意，我们可以优化模型，优化的方法很多，其中网格搜索参数是一种有效方法，当然我们也可以采用手工调节参数等方法。如果出现过拟合，尤其是回归类问题，我们可以考虑正则化方法来降低模型的泛化误差

d.训练模型的工程化和服务化

完成模型的尽力获取理想的效果，因而再额外花时间和精力确定合适的计算资源和架构，试图在不同的生产环境下大规模复现模型并实现相同的效果，是一件需要周密计划的任务。创建API：等模型可以打包入 API 时，就可以根据需求用 Flask创建 API、理想状态下，你必须创建 REST式的 API，因为这样有助于分离客户端和服务器，也能优化可见性、可靠性和可扩展性。要做一个全面的测试，确保模型响应 API 的正确预测。

Web服务器：现在是时候为你创建的 API 测试 web 服务器了。如果你是用 Flask 创建的 API，Gunicorn/Apache/Nginx 是个不错的选择。

前端通过vue和bootstrap编写好web UI，和后台api进行交互。

* 1. 实施技术方案所需的条件

数据分析端

Tensorflow version 1.15.0 or higher with Keras API

Pandas

Numpy

后端：

Flask

Apache or nginx

前端：

Bootstrap

Vue

D3.js

* 1. 存在的主要问题 技术关键

数据探索、特征列选取，模型的调优和防止过拟合等方面

* 1. 预期能够达到的研究目标

情感分析是确定语言是正面，负面还是中立情绪的过程。分析客户的情绪对企业有很多好处。 例如公司可以根据情感过滤客户反馈，以识别他们需要改进的服务。公司可以通过监视客户撰写有关其产品的评论的情绪来轻松管理其在线声誉。

# 课题计划进度表

毕业进度安排：

* 1-2周：学习相关理论，熟悉开发环境
* 3-4周：系统设计
* 5-7周：深度学习算法模块实现
* 8-10周：服务器模块实现
* 11-12周：前端模块实现
* 13-15周：完成毕业论文，提交软件及相关文档
* 1-15周：外文翻译
* 1-15周：毕业论文

# 参考文献

外文翻译要求

外文资料应与课题相关，且在3年以内，内容应相对完整，避免取长论文的部分段落。

**开题答辩用PPT内容安排**

首页    课题名称、姓名、导师

第1页：课题的主要内容

第2页：课题任务

第4～N页：技术方案

第N+1页：存在的主要问题和技术关键

第N+2页：预期目标

第N+3页：研究进度