北 京 邮 电 大 学

本科毕业设计（论文）开题报告

学院	计算机学院	专业	计算机科学与技术	班级	2014211307
学生姓名	王超	学号	2014211310	班内序号	07
指导教师姓名	刘晓鸿	所在单位	计算机学院	职称	副教授
设计（论文）题目	（中文）基于循环神经网络的影评情感分析系统的设计与实现				
	（英文） Design and implementation of the movie-review sentiment classification system based on RNN				

毕业设计（论文）开题报告内容：（主要包含选题的背景和意义；研究的基本内容和拟解决的主要问题；研究方法及措施；研究工作的步骤与进度；主要参考文献等项目）

1. 选题的背景和意义

信息传播的方式经历了语言符号、文字、电子传播等演变过程。21 世纪起，进入了互联网时代，越来越多的用户开始藉由互联网发表着自己的观点，传播个人的信息。任意一项商品或服务就会产生成千上万的评价信息，这就使得用户评论成为了一种越发重要的信息载体。

伴随着互联网的发展，电子商务的领域也逐渐扩散，电影与互联网产业也渐渐跨界融合。电影市场的下游环节，用户通过在线票务系统选择影片、影院并在观影结束后进行点评，这一系列的行为所产生的大量数据，能够为电影产业的上游环节提供重要思考。

在传统电影行业中是“我生产什么观众就看什么”，用户只能被动接受，可选择性小。而在互联网环境下，观众由产生观影意向到走进影院、完成观影，会经过“搜索关注、购票观看、评分评论”三个阶段，不管在哪一个阶段，用户的主体地位都已经被默许。[1]

2012 年，国内学者殷国鹏等[2]以信息采纳理论为研究框架，探讨了消费者在购买决策中采纳与接受在线评论信息的两类影响因素，即评论本身特征和评论者要素，并结合社会网络视角构建了在线评论有用性模型。社交媒体的电影评论对于购票者做出决产生了一定的影响。制片商也能够根据相关的影评来调整营销策略。

但是，面对如此庞大的评论文本信息，传统的通过人工获取评论的情感倾向是一件特别费时费力的事情，因此，如何利用自然语言处理领域的相关技术针对性地对评论文本的情感倾向进行自动化地挖掘与分析成为当今热门且很有意义的研究课题。

2. 研究的基本内容和拟解决的主要问题

2.1 研究的基本内容

主要内容：

1.搭建基于 Tensorflow 的深度学习网络平台，并获取影评数据集；

2.使用不同结构的深度网络模型，对数据集进行训练，得到相应的深度神经网络；

3.通过训练时间和计算时间及网络规模之类特性对比，得到合适的神经网络，并对实验中的结果进行分析解释；

4. 基于上述得到的神经网络实现影评情感分析软件。

2.2 拟解决的主要问题

1) 学习 word2vec 原理及应用，学习 CBOW 与 Skip-Gram 神经网络语言模型，以及 TF-IDF 模型，根据训练性能调整 word2vec 模型。
2) 学习 CNN、RNN、LSTM、GRU 等网络模型， 比对不同结构的深度网络模型，最终选择合适的神经网络
3) 学习过抽样、欠抽样、集成学习等方法，解决数据不均衡问题

3. 研究方法及措施

3.1 获得社交媒体上的电影评论数据；

在具体的研究过程中，通过网络爬虫分别采集了所需的评论数据，数据采集来源为豆瓣网（www.douban.com）。 豆瓣网作为国内最主要的以电影、图书及音乐作为纽带的社交网站，它不仅有关于国内外各种电影的用户评论，并且在用户群体中既有专业影评人，也有普通观影用户。 基于以上原因，本文选择豆瓣网影评作为研究对象，并对其评价数据进行实例分析。

3.2 搭建循环神经网络模型，对电影评论进行情感分析训练和建模；

寻找合适的语料库（如 jieba 语料库等）对数据集进行分词, 去除停用词，完成对数据集的预处理操作。

基于 word2vec 使用 CBOW 或 Skip-Gram 模型得到词向量。[3]考虑到可能 word2vec 无法区分文本中词汇的重要程度，可能进一步引入 TF-IDF 模型计算 word2vec 词向量的权重，提出加权 word2vec 模型。[4]或者通过构建情感要素词典（如 HowNet 的“情感分析用词语集”）捕捉含情感要素的词，通过构建词的情感特征向量来表示这些词的情感要素，然后与 word2vec 词向量进行特征融合，构成多元特征词向量。[5]

对 CNN 模型、传统 RNN 模型、LSTM 等模型进行性能比对，选择合适的神经网络。[6–8]

3.3 处理数据中可能存在的不均衡数据问题，改进得到的模型，提升模型分类准确率和召回率；

对文本分类的数据集中的不均衡问题处理的最终目标是在确保总体的分类准确度不降低或降低不大的前提下，提高少数类的分类精度。[9]

在数据不均衡情况下，模型性能评价指标只采用准确度(accuracy)是不恰当的。所以，使用如混淆矩阵、 F1 得分、ROC 曲线等更有效的评价作为评价标准。

对数据不均衡有下面几类处理方法:

3.3.1 数据集层面的方法

可以采用过抽样如 SMOTE 过采样算法、欠抽样策略等策略进行调整。[10]

3.3.2 算法层面的方法

可以通过改变概率密度、集成算法（Boosting、抽样和集成算法的融合）等方法进行调整。[10]

3.4 将模型应用到其他可能场景，证明模型有效性及其泛化能力。

4. 研究工作的步骤与进度

3 月 5 日---4 月 1 日：搭建环境，读相关文献；
4 月 2 日---4 月 29 日：进行算法设计，比较不同方法的结果，并进行编码和调试；
4 月 30 日---6 月 3 日： 通过不同结果对比，优化设计实现；
6 月 5 日---6 月 22 日：撰写论文及答辩。

5. 主要参考文献

[1] 沈尧. 中国电影在线票务发展研究[D]. 中国电影艺术研究中心, 2016.
[2] 殷国鹏, 刘雯雯, 祝珊. 网络社区在线评论有用性影响模型研究——基于信息采纳与社会网络视角[J]. 图书情报工作, 2012(16): 140–147.
[3] ZHANG L, WANG S, LIU B. Deep Learning for Sentiment Analysis : A Survey[J]. arXiv:1801.07883 [cs, stat], 2018.
[4] 张谦, 高章敏, 刘嘉勇. 基于 Word2vec 的微博短文本分类研究[J]. 信息网络安全, 2017(01): 57–62.
[5] 李科. 基于多元特征融合和 LSTM 神经网络的中文评论情感分析[D]. 太原理工大学, 2017.
[6] KIM Y. Convolutional Neural Networks for Sentence Classification[J]. arXiv:1408.5882 [cs], 2014.
[7] WANG J, YU L-C, LAI K 等. Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model[C]//2016: 225–230.
[8] 李丹. 基于长短时记忆网络的中文文本情感分析[D]. 北京邮电大学, 2017.
[9] 谢娜娜. 基于不均衡数据集的文本分类算法研究[D]. 重庆大学, 2013.
[10] 陶新民, 郝思媛, 张冬雪等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报(自然科学版), 2013(01): 101–110+121.

指导教师签字		日期	年 月 日

注：可根据开题报告的长度加页。