成绩	优秀

毕业设计

题目：基于 EWMA 和 RNTN 的电商评论分析研究与应用

学 生 姓 名：	高文光
学　　号：	1708040105
所 在 学 院：	计算机学院
专 业 名 称：	软件工程
届　　别：	2021 届
指 导 教 师：	孔军
职称（学位）：	助教(硕士)

淮南师范学院教务处制

淮南师范学院本科毕业论文（设计）
诚信承诺书

1. 本人郑重承诺：所呈交的毕业论文（设计），题目《　　》是本人在指导教师指导下独立完成的，没有弄虚作假，没有抄袭、剽窃别人的内容；

2. 毕业论文（设计）所使用的相关资料、数据、观点等均真实可靠，文中所有引用的他人观点、材料、数据、图表均已注释说明来源；

3. 毕业论文（设计）中无抄袭、剽窃或不正当引用他人学术观点、思想和学术成果，伪造、篡改数据的情况；

4. 本人已被告知并清楚：学院对毕业论文（设计）中的抄袭、剽窃、弄虚作假等违反学术规范的行为将严肃处理，并可能导致毕业论文（设计）成绩不合格，无法正常毕业、取消学士学位资格或注销并追回已发放的毕业证书、学士学位证书等严重后果；

5. 若在省教育厅、学院组织的毕业论文（设计）检查、评比中，被发现有抄袭、剽窃、弄虚作假等违反学术规范的行为，本人愿意接受学院按有关规定给予的处理，并承担相应责任。

学生（签名）：

日期：　　年　　月　　日

目　　录

基于 EWMA 和 RNTN 的电商评论分析研究与应用

学生：高文光（指导老师：孔军）

（淮南师范学院计算机学院）

摘 要：本文基于亚马逊网站提供的数据，对阳光公司三种产品的文本评论进行分析，提取了用户关注的产品关键特征。通过多元线性回归模型和 EWMA 方法得到在线市场的成功因素和声誉趋势，并利用 logistic 函数模型显示产品成功概率，最后基于 RNTN(递归神经张量网络模型)将评论分成非常消极、消极、中立、积极和非常积极五种类别，使用皮尔逊和斯皮尔曼相关性分析对评论类别向量和评级向量进行相关性分析。

关键词：评论；模型；相似度；机器学习；情感分析

Research and application of e-commerce review analysis based on EWMA and RNTN

Student: GAO Wenguang(Faculty Adviser：KONG Jun)

(School of Computer Science, Huainan Normal University)

Abstract: Based on the data provided by amazon.com, this paper analyzes the text reviews of three products of sunshine company, and extracts the key features of products concerned by users. Through multiple linear regression model and EWMA method, we get the success factors and reputation trend of online market, and use logistic function model to show the product success probability, Finally, based on RNTN (recurrent neural tensor network model), reviews are divided into five categories: very negative, negative, neutral, positive and very positive, Pearson and Spearman correlation analysis are used to analyze the correlation between the comment category vector and the rating vector.

Keywords: review; model; similarity; machine learning; sentiment analysis

前言

随着社会迅速发展，我们已经进入一个互联网的智能新时代，我国人民的经济水平也在提高。不少人开始追求物质上的满足，从以前买东西都要去集市，到现在足不出户就能买到物美价廉的物品，人民的生活从而方便了起来，网上交易逐渐占据主流。因此对于大多数电商来说，如何通过用户的产品评论中来提取用户需求是一个急需解决的问题。

本文综合考虑了亚马逊三类产品数据集中有意义的评论,并将评论与星级量化,分别抽取了若干特征值较高的指标,构建回归模型,以此来给出三类在线产品是否取得成功的具体描述。在综合评价的基础上我们采用了EWMA算法来表明产品在在线市场中声誉的上升和下降趋势。对于三种产品在线出售后,我们构建了逻辑回归模型来评判三种产品的成功和概率。对于特定星级的评论,我们考虑了两周的评论星级分值,运用相关性分析来确定客户是否会因之前的评论撰写更多可能的评论。对于一些“热情”,和“失望”之类的描述评论,判定是否与其它相关得分紧密相连。

本文的优点在于建立了EWMA模型，可以在随机性的评论中获得可靠、平滑的数据,能够较好的分析趋势,并且通过移动加权可以模拟现实中的评分。缺点在于问题一人为给评论特征赋值,主观性较强。

1 绪论

1.1 研究目的意义

现如今，随着越来越多的人开始网上购物和互联网的广泛普及，电商如雨后春笋般成立，在其创建的线上商城中为客户提供了对交易进行评分和评价的机会，网络上出现了大量主观性电商评论。这些电商评论不仅包括客户对所购买产品的满意度和意见，也包括其他客户在这些评论中提交有帮助或无帮助的评分。各电商公司使用这些电商评论来深入了解其参与经营的市场，通过大量主观性评论来分析客户对产品的需求，以便优化产品质量及在线销售方式，用于增强潜在客户对产品的满意度。

1.2 国内外研究状况和应用背景

近年来，由于全球化网络的快速发展，互联网进入大众生活各个领域，因而产生了一种通过网络进行物品交易的方式，大大促进电商如雨后春笋般增长。典型性的代表有“淘宝”、“亚马逊”、“京东”等。

据网经社“电数宝”数据库显示，中国在2017年至2018年期间，通过网络电商交易的金额达到几十万亿元，同时较去年电商交易金额增长13.5。同时对于全球而言，电商极大加快了经济增长，丰富了物品交易的方式且极大促进了巨大的交易规模。并且由于电商的产生，包括亚马孙、阿里巴巴、京东和Shopify等企业巨头都在投入巨资进军互联网行业。截止到2018年为止，全球电商服务直接从业人员超过上千万人，由电商服务间接带动的各种领域就业人数已超过一亿人。在直接就业上，随着互联网领域和电子商务领域的不断深入和发展，电商服务从业人员的规模也呈现出不断攀升的趋势。线上与线下相结合的交易方式越发频繁，从业人员的规模也随之得以快速增长。在间接就业上，在最近几年随着大量短视频电商的火热，一些网红电商行业、直播电商行业等新兴团体的接连出现，诞生了很多新的服务岗位与就业机会。电商传播方式从最初静态的图片到之后动态的视频，背后的是流量的转移和获取。

1.3 研究内容和方法

1.3.1 研究内容

近些年来大数据与网络的兴起,可以在商业、经济和其他领域范围内，将信息越来越多地以数据和分析为基础,而不是以经验和直觉为基础,能够及时将数据量化处理进行综合评分,然后得到有用的信息,来帮助我们在实际问题中更好的处理。

亚马逊在创建的线上商城中为客户提供了进行评分和评级的模式,阳光公司根据所销售的三种产品的评级和评论来识别关键模式、度量和参数,来进行销售方式的改进并识别潜在的重要设计功能,以增强客户对产品的满意度。

其主要研究内容与创新如下：

（1）根据分析提供的三个产品数据集，我们使用了TF-IDF（词频-逆向文件频率）算法和多元线性回归分析方法，建立了多元线性回归模型，得出三类产品的特征与星级评分之间关联度大小的结论。首先，我们对数据集进行清洗，然后用TF-IDF算法找出对应的关键词，并从数据集中筛选出较强特征值的指标，最后构建星级评分与指标评分之间的多元线性回归模型，得出二者之间关系，并提出针对性建议。

（2）根据数据集的评论和评级，我们采用余弦相似度算法，建立了综合评价模型，得出我们运用此方法的合理性。首先运用TF-IDF将具有代表意义的评论分类量化，然后利用余弦相似度算法计算评论与我们所建立的评论类别之间的相似度来给用户评论进行打分，最后通过相似度得分来描述最佳测量度。

（3）根据在每个数据集的数据中辨别并发现基于时间的度量和模式，我们采用移动加权平均算法,建立了EWMA模型，使得我们能够得到更加可靠的数据。通过对数据进行加权移动平均,然后得到平滑的数据，并对数据进行拟合，根据所得趋势图可以发现吹风机、奶嘴和微波炉的声誉趋势。

（4）根据基于文本和评级的度量的组合,我们采用了logistic回归方法,建立了logistic回归函数模型,得出三类产品的概率值。首先将三类产品评价得分带入逻辑回归函数训练,然后分别得出吹风机,奶嘴,微波炉概率值,最后得出奶嘴、微波炉与吹风机这三种产品是否相对成功。

（5）根据各个产品的数据集，我们采用皮尔逊相关分析，建立了典型相关性分析模型得出两者之间相关性得分。首先,我们对两周的数据预处理,然后得到两周的评论与星级的综合得分，将分数进行训练，最后得出前一周的得分与最新一周得分相关性大小，推测客户会有可能撰写某种类型评论的结论。

（6）根据数据集的评论描述，我们采用RNTN方法，通过构造递归神经张量网络模型分析评论内容，得出评论等级为-2~2的区间范围。首先,我们对一个月的数据预处理,然后将得分进行训练，将评论分成五种类别，采用斯皮尔曼相关性对评论类别向量和评级向量进行相关性分析，得出评论类别和评级之间相关性大小。

1.3.2 研究方法

本文基于亚马逊网站提供的数据，通过多元线性回归模型和EWMA方法，并利用logistic函数模型，最后基于RNTN方法将评论分成五种类别，使用皮尔逊和斯皮尔曼相关性分析对评论类别向量和评级向量进行相关性分析。

2 基于电商评论研究综述

2.1 研究内容分析

（1）首先，一般用TF-IDF(词频-逆向文件频率)算法和多元线性回归分析方法。对所给的数据特点进行分析后,我们采用了TF-IDF算法对数据进行预处理工作。我们得出此问题具有评价与回归类问题的特性,因此我们基于线性回归理论建立了多元线性回归数学模型,然后根据这个模型我们对结果分别进行分析,并将结果进行比较,得出了三类在线产品的星级评分与各自特征值之间的关系,从而描述了在线市场产品的成功因素。

（2）通过研究分析,对于解决文中最佳测量度问题一般用TF-IDF（词频-逆向文件频率）算法与余弦相似度算法分析。对数据清洗后数据特点进行分析，我们首先运用TF-IDF算法将具有代表意义的评论分类量化,然后利用余弦相似度算法计算评论与我们所建立的评论类别之间的相似度,最后通过相似度得分来描述最佳测量度。

（3）对于如何判定声誉趋势,我们可以判定属于一个时序分析类问题,而解决这个问题需要用到时间序列方面的理论。对于解决此类问题我们可以用EWMA（指数加权移动平均法 ）算法分析。对数据清洗后数据特点进行分析,我们得出此问题具有指数加权类问题的特性,因此我们基于指数加权理论建立了EWMA数学模型,然后根据这个模型我们对结果分别进行分析,并将结果进行比较,得出了产品在在线市场中声誉的上升和下降趋势。

（4）对于判定产品成功与概率,我们可以判定属于逻辑回归类问题,而解决这个问题需要用到回归分析方面的理论。对数据清洗后数据特点进行分析,我们得出此问题具有回归类问题的特性,因此我们可以基于回归理论建立了逻辑回归数学模型,然后根据这个模型,我们得到回归后的概率值,并对结果分别进行分析,并将结果进行比较,得出三类产品的成功概率。

（5）对于特定星级是否会影响客户评论,我们可以判定属于典型相关性问题,而解决这个问题需要用到相关性分析方面的理论。因此我们基于典型相关性分析理论分析了我们所采样的数据,得出了分析后的关联程度值,并将结果进行分析比较,得出了我们所假定

的时间节点前后的星级评论分数之间的关联程度。

（6）对于情感词汇是否与其他相关得分相联，我们可以通过构造RNTN递归神经张量网络模型分析评论内容，得出评论等级为-2~2的区间范围。对清洗后数据特点进行分析,我们得出此问题具有相关性问题的特性,然后将得分进行训练，将评论分成五种类别，采用斯皮尔曼相关性对评论类别向量和评级向量进行相关性分析，得出评论类别和评级之间相关性程度。

2.2 模型假设

（1）三种商品数据都不再区分子商品。在三种数据中我们均发现，子产品数量较少，同时由于数据量的限制，可以忽略子产品带来的影响，所以将三种商品数据看作成三种独立的商品，不再区分子商品。

（2）忽略快递运输问题。由于快递运输过程中产生的一系列情况不能为阳光公司的三种商品改进做出决策。

（3）模型星级与评论正相关。分析三种商品的数据，同时结合现实中对商品的评价出现星级高，评论差，星级低，评论好的概率较低的情况可得。

2.3 符号说明

表 1　符号说明

符号	含义
ε	误差扰动随机项
$v2$	加权下降的速率
$I1$	吹风机星级变量
$I2$	微波炉星级变量
$I3$	奶嘴星级变量

2.4 数据预处理

2.4.1 数据展示

对于亚马逊三类产品数据集，每一个产品数据集都有相同的特征指标，具体如下显示（图 1）。

marketplac	customer_i	review_id	product_ic	product_pa	product_ti	product_ca	star_ratir	helpful_vc	total_vote	vine	verified_p	review_hea	review_bo	review_date
US	34678741	R9T1FE2ZX2	B003V264W7	732252283	remington	Beauty	5	0	0	N	Y	Works gre	Works gre	8/31/2015
US	11599505	RE36JAD5V8	B0009XH6V4	670161917	andis mic	Beauty	4	0	0	N	Y	I love tra	This dries	8/31/2015
US	2282190	RIDHM8B7S(	B0007NZPY6	16483457	conair pr	Beauty	5	0	1	N	Y	Five Stars	Love this	8/31/2015
US	43669858	R14QGWPCHU	B00BB8ZIW(	253917972	remington	Beauty	5	0	0	N	Y	Five Stars	styling ha	8/31/2015
US	107098	R35BHQJHX)	B003V264W7	732252283	remington	Beauty	4	0	0	N	N	I think's	I just got	8/31/2015
US	51995766	R230LCPQD(	B000065DJ)	919751065	revlon 18	Beauty	5	0	0	N	Y	Five Stars	Excellent	8/31/2015
US	39431051	R21NN9ONV2	B000FS1W4(	235105995	revlon ess	Beauty	1	0	0	N	N	Gets extre	Gets extre	8/31/2015
US	180659	RYOOYLVIAI	B003FBGS8I	195677102	conair pr	Beauty	3	1	1	N	Y	Everything	I found ev	8/31/2015
US	17023782	R18NK8BQ5I	B0057HQ6C2	582752797	pibbs tte	Beauty	5	0	0	N	Y	Five Stars	Perfect	8/31/2015
US	17563775	RDOBGSERM2	B00132ZG3(	758099411	conair 18	Beauty	5	0	1	N	Y	Nice hair	I really l	8/31/2015
US	181650	R2NEOR5Y0I	B003FBGS8I	195677102	conair pr	Beauty	5	0	0	N	N	Awesome d	Got tho or	8/30/2015
US	9924936	R3N0F2FKJ(	B00CC7Y0G4	253762851	remington	Beauty	5	3	3	N	Y	Works gre	Reckon I l	8/30/2015
US	41150214	R2M6U66CT3	B00092M2X2	221722169	hot tools	Beauty	3	0	0	N	Y	The stylir	The stylir	8/30/2015
US	49876147	R2WDYCZI9)	B00006IV2:	357308868	conair 18	Beauty	4	0	0	N	Y	Blows like	Nice hair	8/30/2015
US	42278060	R1DZKVY1C)	B00A7JA72(	646926938	babyliss r	Beauty	2	1	1	N	Y	Great drie	Great proc	8/30/2015
US	47929663	R3IL7ENAGV	B001BOFIR(	26711891	andis tou	Beauty	5	0	0	N	Y	A good re	Like this	8/30/2015
US	15822780	R1PFMP0SD2	B00SKQFT4(	593915883	xtava all	Beauty	5	0	1	N	Y	Girlfrien	Very quiet	8/30/2015

图 1 数据各类指标

由图1得知，主要指标有star_rating、helpful_votes、total_votes、vine、verified_purchase、review_headline、review_body和review_date。

表 2　主要指标释义

指标	释义
star_rating	星级
helpful_votes	有帮助的投票
total_votes	总投票
vine	是否为可信度高的用户
verified_purchase	是否购买过
review_headline	评论标题
review_body	评论内容
review_date	评论日期

2.4.2 数据筛选与分析

（1） 针对vine和verified_purchase要求说明,对vine =N和verified_purchas = N进行数据筛选。

（2）为了方便接下来的数据处理和理解，需要进行数据格式统一，处理大小写问题和把N和Y进行量化成0-1。

（3）根据对数据进行描述性统计，我们得到数据筛选后评论的数量为9989条，商品的类别为424种，积极、中立和消极的百分比分别为78.68%、8.80%和12.52%。

2.4.3 数据可视化

2.4.3.1 星级和 Vine 分布

通过星级和Vine分布图显示（图2），我们得出了星级个数分布情况，其中5星级最多，2星级最少。

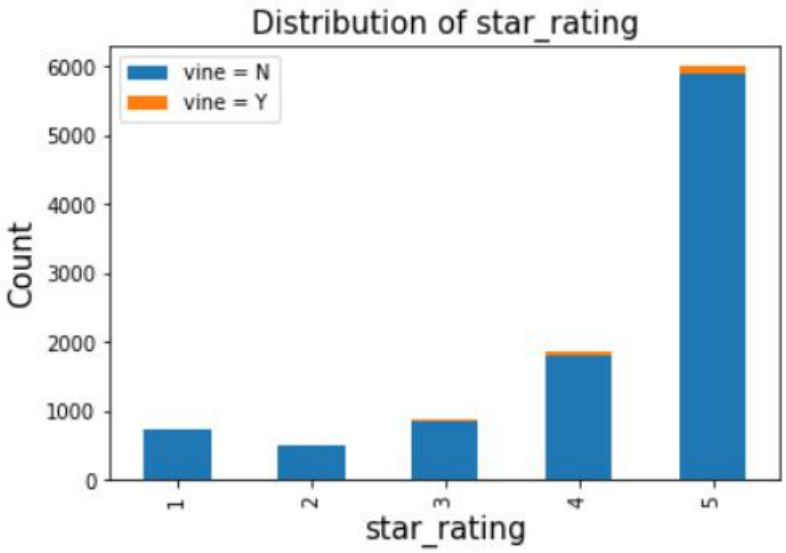

图 2 星级和 Vine 分布

2.4.3.2 评论观点态度分布

通过评论观点态度分布图（图3），我们得出观点分布以及报告分布情况，其中积极占比最大消极与中立占比较少，由此我们得出大多数客户对产品比较满意。

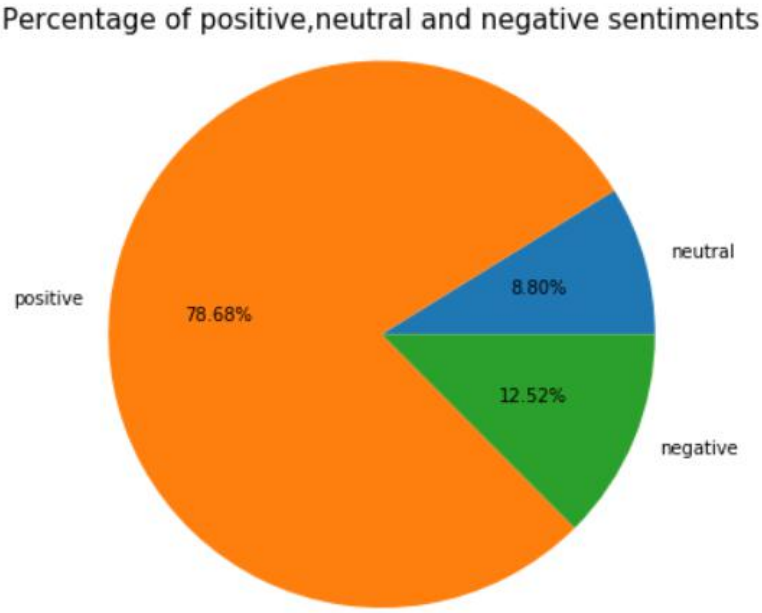

图 3 积极、中立和消极百分比

2.4.4 数据清洗

数据清洗过程显示（图4），首先要对数据进行数据预处理，我们通过所处理后的数据集,将数据集评论里明显无意义的词汇进行汇总生成停止词文件。使用jieba模块实现简单的分词,由分词结果返回一个生成器,自定义简单分词函数,通过Apply函数形式分词生成评论语料，加载停止词,并转化成词列表,读出评论语料,通过词列表的处理，转化成

单词矩阵。将单词矩阵通过gensim中的Dictionary模块生成字典词库，构建词和编号的映射关系,生成BOW稀疏矩阵，将评论语料转化成稀疏矩阵。最后通过TF-IDF算法转化成TF-IDF格式，通过得出的结果我们找出评论中重要的词，再次屏蔽全局评论出现频率较高的无意义词，构建LDA模型[3],输出评论中出现频率最高的前n个词，数据清洗完成。

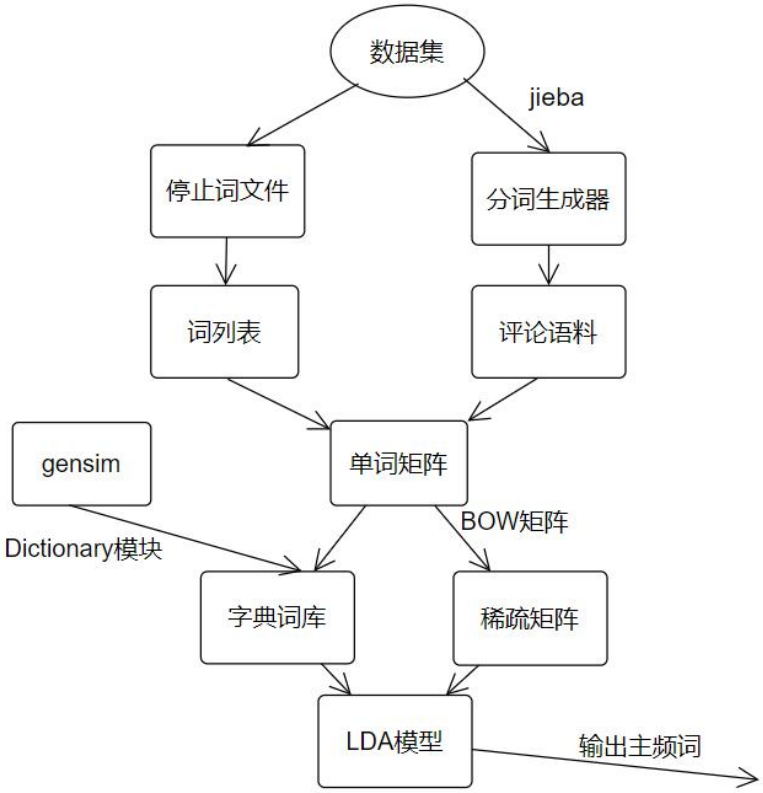

图 4 数据清洗流程

3 基于电商评论研究模型建立

3.1 基于多元线性回归建立多元线性回归模型

3.1.1 模型的结构分析

多元线性回归模型的一般表达式:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... \beta nXn + \varepsilon \quad (1)$$

其中βi (i=1,2,⋯,n)为多元线性回归系数,Y为被解释变量，Xj (j=1,2,⋯,n)为对Y的解释变量， ε 反映的是误差扰动随机项。

3.1.2 模型的分析和建立

考虑到指标评分与星级评分之间的关系,我们运用了TF-IDF算法[3-4]结合回归分析方法构建了多元线性回归模型来模拟星级评分与指标评分之间的关系,当自变量指标评分每上升或下降一个或若干个单位时,都能表示出因变量星级评分的变化趋势,然后基于此变化趋势,我们得出各类指标与销售产品的成功变化率之间的关系,所以我们得出此模型对于处理此类问题有着较高的说服性与实用性,因此选择该模型。

针对此模型我们对指标与星级之间构建模型，从婴儿奶嘴,微波炉,吹风机三类数集中分别抽取前一个月数据集,从中选取具有代表意义的评论,并找出特征指标。

3.1.3 模型的实现与运行结果

```
data = xlsread('hair_dryer.xlsx');
.....
X = [ones(99,1) x1 x2 x3 x4 x5 x6 x7];
[b,bint,r,rint,stats]=regress(y,X);
```

我们对前一个月的数据进行回归拟合,其中

吹风机指标有：A1(体积是否过小),A2(热度是否过热),A3 (速度是否适合),A4(噪声是否小),A5(价格是否实惠),A6(质量是否好),A7(按钮是否方便)；

微波炉指标有：B1(定价是否适宜),B2(体积是否合适),B3 (是否易清洗),B4(款式是否好看)；

奶嘴指标有：C1(价格是否实惠),C2(质量好坏),C3 (外观是否可爱)，C4(样式种类是否多)，C5(是否实用是否方便)，C6(运输是否迅速)；

根据上述指标我们构建三个多元线性回归方程:

$$Y1 = 0.5818A1 + 0.3330A2 + 0.7509A3 + 0.8037A4 + 0.6212A5 + 1.4200A6 +$$

0.1072A7 + 3.0305 （2）

Y2 = 0.5460B1 + 0.5748B2 + 0.4001B3 + 1.1575B4 + 2.7507 （3）

Y3 = 0.436C1 + 0.9090C2 + 0.3836C3 + 0.1796C4 + 0.3279C5 + 0.246C6 + 3.2928（4）

对于吹风机，奶嘴，微波炉，我们多元回归模型拟合出stats的判决系数分别为0.6401、06096、0.6055，由此我们得知对于本模型三类拟合度良好。

由运行结果数据可知，三类检验的统计值分别为:23.1165,24.2174,36.0644,由此我们得知模型整体上解释变量与被解释变量之间线性关系显著。

3.2 基于余弦相似度算法建立综合评价模型

3.2.1 模型的结构分析

余弦相似度[8]计算公式如下:

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{5}$$

余弦相似度算法主要是在一个向量空间中用两个向量间有夹角的余弦值做为衡量个体间的偏移差异最大大小,当余弦的值越来越接近1,就可以说明其夹角越趋于0度,即两个向量越相似,故称之为"余弦相似性"。

3.2.2 模型的分析和建立

考虑到对较强特征值的评论进行向量化分类,我们采用余弦相似度算法,将分类过的特征值评论与一个月的评论进行相似度对比,通过我们所分类评论的得分给一个月的评论进行打分,由此我们就可以根据评论与评级来确定阳光公司的最佳测量度,根据我们所测得的数据,我们得出该模型对于处理此类问题有着较高的说服性与参考性,因此我们选择该模型。

3.2.3 模型的实现与运行结果

```
from sklearn.feature_extraction.text import CountVectorizer, TfidfVectorizer, TfidfTransformer
from sklearn.metrics.pairwise import cosine_similarity
corpus = list(review_2000[::1])
tfidf_vec = TfidfVectorizer()
```

```
tfidf_matrix = tfidf_vec.fit_transform(corpus)
cosine_similarity = cosine_similarity(tfidf_matrix)
df_cosine_similarity=pd.DataFrame(cosine_similarity[[8,11,56,58,61,62,78,1027]])
df_cosine_similarity = df_cosine_similarity.iloc[:,0:1957]
```

导入sklearn中所需模块，将评论载入语料库，通过TF-IDF向量化工具计算分类过的特征值评论与一个月的评论之间的相似度。将分类的特征值进行分类和设置类别分数，通过对比每一次相似度取最大值，对所有的评论进行分类（图5）。

	star_rating	review	review_kind	review_score
0	5	Works great Works great!	11	6
1	4	I love travel blow dryers because they are eas...	68	5
2	5	Five Stars Love this dryer!	68	5
3	5	Five Stars styling hair in style	64	3
4	5	Five Stars Excellent dryer.	68	5
5	3	Everything okay but.....!! I found everything ...	68	5
6	5	Five Stars Perfect	68	5
7	5	Nice hairdryer that works very well. I really ...	67	1

图 5 评论分类

3.3 基于移动加权平均算法建立 EWMA 模型

3.3.1 模型的结构分析

EWMA(指数加权移动平均法)[9-10]公式如下:

$$EWMA(t) = \lambda Y(t) + (1-\lambda)EWMA(t-1) \quad （t = 1, 2, ..., n） \tag{6}$$

其中EWMA(t)为时刻的估计值;Y(t)为时刻的测量值;n所观察的总时间; λ ($0 > \lambda > 1$)表示对于历史测量值的权重系数。

EWMA(指数加权移动平均法)算法是指各数值的加权系数随时间呈指数式递减,如果越靠近当前时刻的数值,加权系数就越大,因此可以很好反映出短期内时刻之间的关联度,便于观察。

3.3.2 模型的分析和建立

考虑到我们所得分的总体趋势,为了更好的显示出总体趋势随时间的变化关系,我们采用EWMA（指数加权移动平均法）算法,目的在与使得我们的数据更加的平滑,并且当我们选取某一固定时间点时,该算法能够准确的显示出我们所选取的时间点与相邻时间点的关联程度，因此我们基于该算法的理论知识,构建出EWMA数学模型,并且根据我们最后所得出的图形,我们能够形象直观的看出变化趋势,得出该模型具有较强的可信度,因

此我们选取该模型。

3.3.3 模型的实现与运行结果

```
def EWMA(scores, alpha=0.9):
    newScores = np.zeros(shape = scores.shape)
    newScores[0] = scores[0]
    for i in range(1, len(scores)):
        newScores[i] = scores[i]*alpha + newScores[i-1]*(1-alpha)
    return newScores;
```

我们选取了一个月数据的总星级和平均星级的加权得分,进行差分。

根据吹风机EWMA曲线图（图6）显示，我们得到吹风机一个月的曲线呈上下波动趋势，在第10至25天波动较大，虽然有部分上升趋势，但总体呈波动趋势，所以我们得出吹风机声誉总体呈不变的趋势。

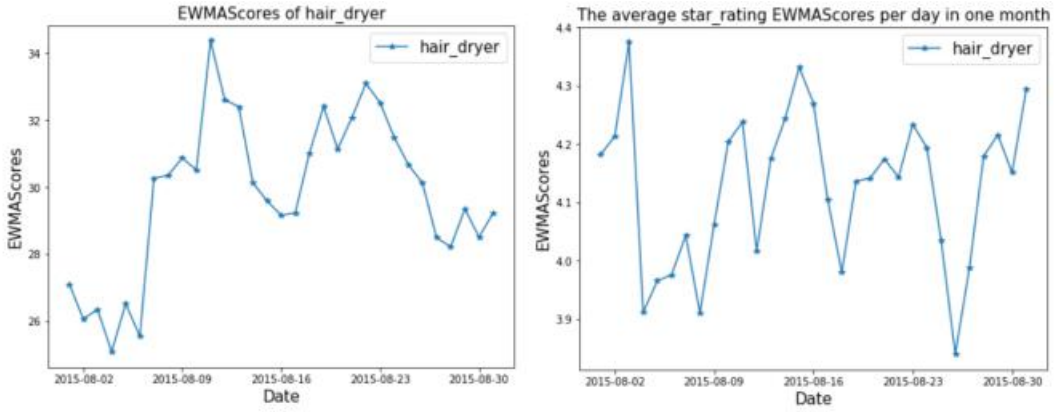

图 6 吹风机 EWMA 曲线

根据奶嘴EWMA曲线图（图7）显示，奶嘴的综合评分经过EWMA模型处理后的分数整体减小，同时对相同时间内每天的平均星级分数进行EWMA模型处理，得到的EWMAScores整体减少，证明奶嘴在在线市场中的声誉呈下降趋势

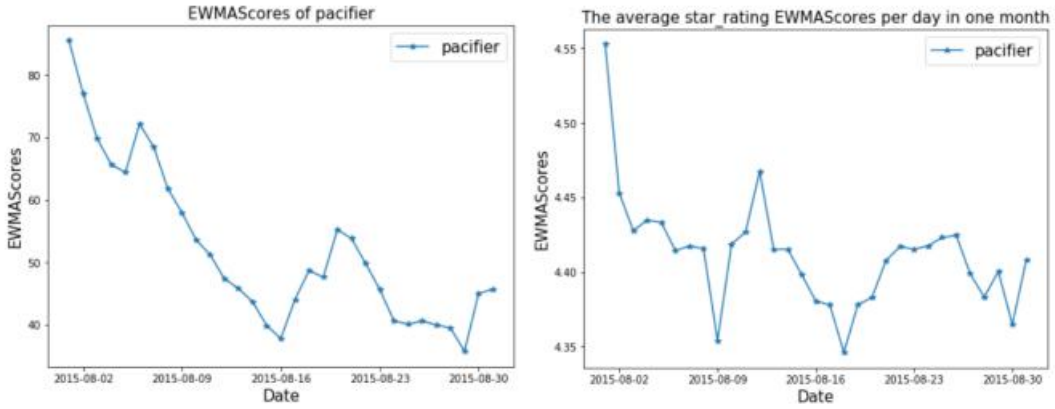

图 7 奶嘴 EWMA 曲线

根据微波炉EWMA曲线图（图8）显示，我们得到微波炉一个月的曲线呈下降趋势，在第5至10天和第20至25天虽然有部分上升趋势，但总体呈下降趋势，所以我们得出微波炉声誉在下降。

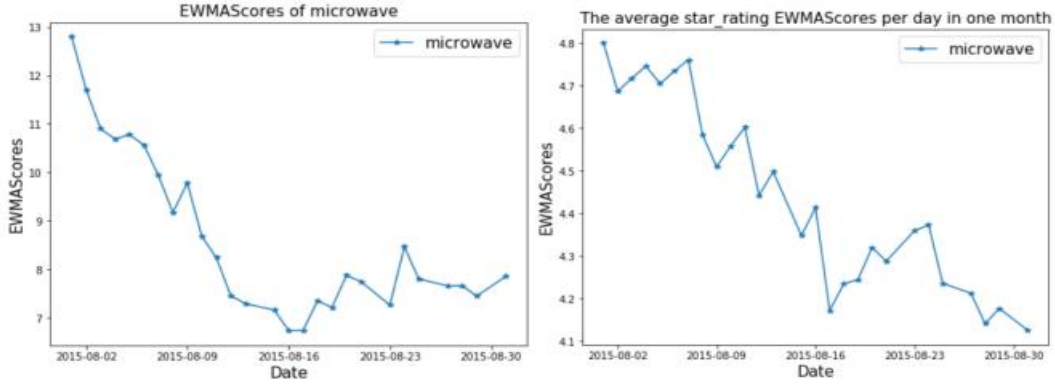

图 8 微波炉 EWMA 曲线

3.4 基于 logistic 回归方法建立逻辑回归函数模型

3.4.1 模型的结构分析

逻辑回归模型是用来研究某些现象发生的概率的模型,一般研究定性的变量,比如电商的成功与失败，是一种二分类模型。

逻辑回归函数模型公式如下:

$$f(x)=\frac{1}{1+e^{-x}} \tag{7}$$

3.4.2 模型的分析和建立

对于产品的潜在成功或失败的度量方法，我们考虑到成功与失败为一个二分类问

题，我们假定成功为1,失败为0。因此我们考虑运用逻辑回归模型来评判产品成功或者失败的概率,如果概率值大于某一个常数,则认为是成功的,如果概率值小于某一个常数,则认为是失败的,由此我们就能很好的度量产品的潜在成功与失败。

Logistic函数，输出0-1内的数，映射成概率值。公式导入代表的是一天的综合评价得分,当趋于无穷大时,可以看作趋于1,而当趋于无穷小时,可以看作趋于0,我们计算出每一天和前一天的得分差值,并将评论数作为权重，差值作为当天得分,累加求和,将求得的结果放入函数中，得出的概率值,如果获得的概率值超过0.5,则认为产品成功,小于0.5,则认为产品失败。

3.4.3 模型的实现与运行结果

```
cumsum = 0
for i in range(0,30):
    cumsum1 = sum_score['total_score'][i+1] - sum_score['total_score'][i]
    cumsum = cumsum + cumsum1
print('cumsum=',cumsum/30)
def sigmoid(z):
    return 1.0/(1 + np.exp(-z))
print('吹风机概率值为',sigmoid(cumsum/30))
```

我们分别将三种产品的数据集导入函数中进行训练，得出三种产品的概率值如下:

奶嘴概率值为0.244；微波炉概率值为0.487；吹风机概率值为0.570。因此我们得出奶嘴与微波炉产品并不成功,吹风机产品成功。

3.5 基于皮尔逊相关系数建立相关性分析模型

3.5.1 模型的结构分析

皮尔逊相关系数是用于度量两个变量X和Y之间的相关程度的量。

皮尔逊相关系数定义公式:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} = \frac{E((X-EX)(Y-EY))}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{8}$$

公式是由变量X和Y之间的协方差和标准差的商来定义的。

3.5.2 模型的分析和建立

考虑到特定的星级对于评论数量的影响，其中特定的评论包括高星级评论与低星级评论，我们选取两个时间段的评论的综合评分，进行皮尔逊相关性分析，检测所得结果是否具有合理性，如果具有合理性，则该模型成立。

3.5.3 模型的实现与运行结果

```
R=np.array([X,Y])
R_mean= np.mean(R,axis=0)
R_std = np.std(R,axis=0)
R_zscore= (R-R_mean)/R_std
R_pd = pd.DataFrame(R_zscore.T, columns=['c1', 'c2'])
R_corr= R_pd.corr(method='spearman')
```

首先,我们对两周的数据预处理,然后得到两周的评论与星级的综合得分，将分数进行训练，最后得出三类产品的相关性：微波炉：0.875，吹风机:0.797，奶嘴:0.715。

结果分析：根据以上分析，我们得出三类产品的评论与星级的相关性较好，显著性较高，所以我们得出两周评论与星级显著性相关，所以特定评论会对其它时间段评论有影响。

3.6 基于 RNTN 建立递归神经张量网络模型

3.6.1 模型的结构分析

递归神经张量网络[14]模型的核心由一个 LSTM 单元组成，可以在某一时刻分析一个词语或者计算语句可能的延续性的概率，并且可以用词向量来组成一个解析数的形式，其初始网络存储状态为一个零矢量并在接下来读取每一个词语后更新直至句尾。

从RNTN分析积极评论图（图9）可以看出，分析该评论得到情感词汇great为非常积极，该评论明显表现对商品非常满意。

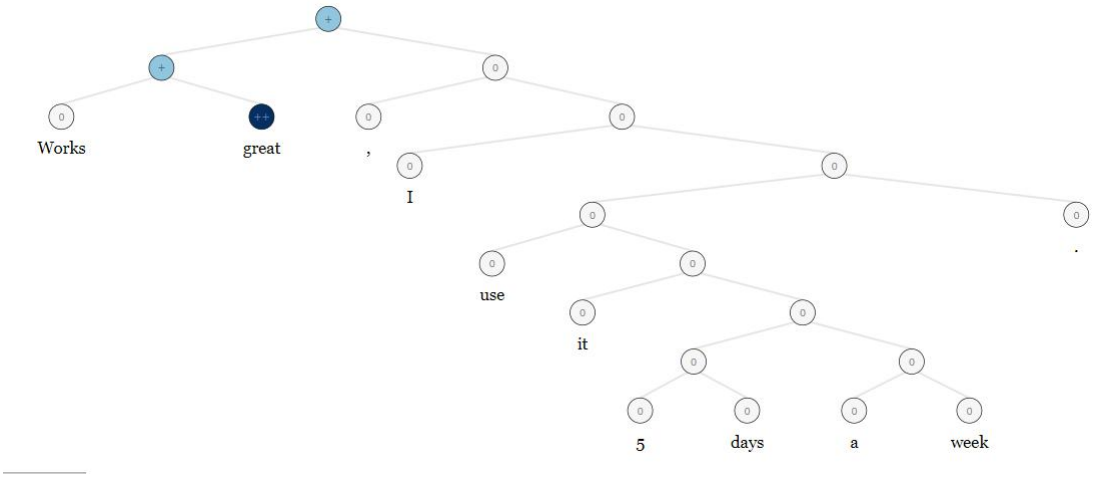

图 9 RNTN 分析积极评论

从RNTN分析消极评论图（图10）可以看出，分析该评论得到情感词汇annoying为非常消极，该评论明显表现对商品非常失望。

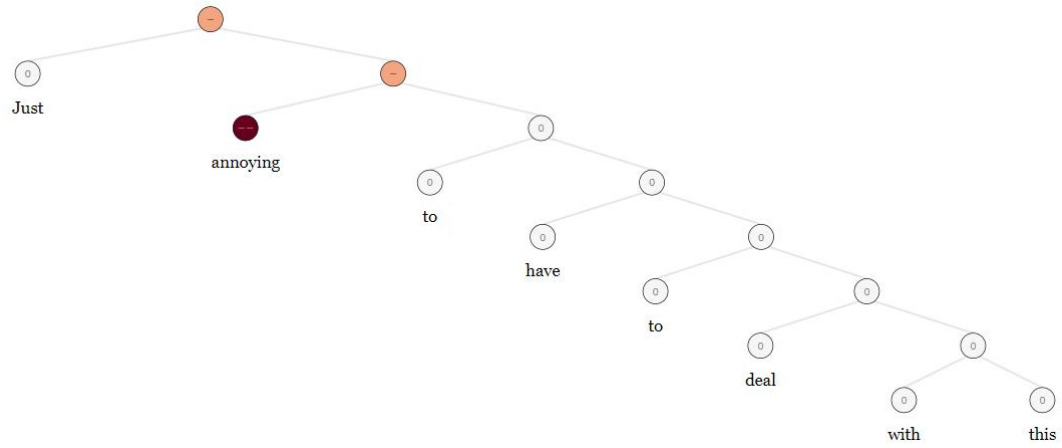

图 10 RNTN 分析消极评论

3.6.2 模型的分析和建立

考虑到评论中的特定的一些评论中如“热情”，“失望”的这些词语，是否会对其它得分产生影响，我们对于此类问题的方法是用RNTN方法，通过构造递归神经张量网络模型模型分析评论内容，得出评论等级为-2~2的区间范围。首先,我们对一个月的数据预处理,然后将得分进行训练，将评论分成非常消极、消极、中立、积极、非常积极五种类别，并使用斯皮尔曼相关性对评论类别向量和评级向量进行相关性分析，最后得出评论类别和评级之间相关性显著的结论。

3.6.3 模型的实现与运行结果

```
import tree as treeM
train = treeM.loadTrees()
```

```
numW = len(treeM.loadWordMap())
rntn = RNN(10,5,numW,mbSize=4)
rntn.initParams()
mbData = train[:1]
rntn.check_grad(mbData)
```

根据RNTN模型找出了-2~2的评论等级区间范围，他们依次对应着非常消极、消极、中立、积极、非常积极五种类别，然后我们将处理好的数据使用斯皮尔曼相关性分析，对评论类别向量和评级向量进行相关性分析，最后得出以下相关系数：奶嘴:0.859；吹风机:0.887；微波炉：0.826

结果分析：根据以上结果我们得出三类产品评论类别与评级之间相关性较好，说明我们对此评论中的特定的一些评论中如“热情”，“失望”的这些词语和其它得分紧密相联。

4 模型小结和改进

4.1 假设

（1）三种商品数据都不再区分子商品。在三种数据中我们均发现，子产品数量较少，同时由于数据量的限制，可以忽略子产品带来的影响，所以将三种商品数据看作成三种独立的商品，不再区分子商品。

（2）忽略快递运输问题。由于快递运输过程中产生的一系列情况不能为阳光公司的三种商品改进做出决策。

（3）模型星级与评论正相关。分析三种商品的数据，同时结合现实中对商品的评价出现星级高，评论差，星级低，评论好的概率较低的情况可得。

4.2 敏感性分析

为了测试分析对最佳测量度的敏感性，我们使用xx商品基于我们模型二建立的模型进行处理得出综合评分，将综合评分使用模型三的EWMA模型得到新的评分为EWMAScores，通过对相同时间内每天的平均星级分数进行EWMA模型处理。从EWMAScores和平均星级基于时间的趋势，我们可以看出两者随时间变化为整体增加，说明该商品在在线市场中的声誉呈上升趋势。

4.3 模型改进

（1）模型四可以通过我们模型二得到的综合评价，建立一个时间序列模型，预测未来15天的综合评价的趋势，如果整体上升则表明该产品是潜在成功的，如果整体下降则表明是潜在失败的，如果波动就是表明维持稳定。

（2）模型三在EWMA模型基础上进行拟合，可以更加直观的表现出趋势。

4.4 优缺点

4.4.1 优点

（1）通过EWMA模型可以一定随机性的评论中，可以获得更可靠的，更平滑的数据与趋势。

（2）通过移动加权可以更好模拟现实中的评分。

4.4.2 缺点

抽取100条样本，给评论特征赋值主观性比较强，选取的典型评论具有一定主观性。

5 总结和展望

本文基于亚马逊网站提供的数据，对阳光公司三种产品的文本评论进行分析，我们提取了用户关注的产品关键特征。通过多元线性回归模型，确定关键特征和评级的关系模式，从而指导产品进行改进和优化。我们为了更好的追踪三种产品上市后的销售效果，采用对用户评论文本进行量化打分的方式，并结合用户评级以及评论点赞数量，来计算当日总分。因为产品的当天得分具有一定的随机性，并且不能完全代表产品的真实销售状态，因此本文采用EWMA方法，将当日得分和历史得分进行加权移动平均，从而获得更加可靠、平滑的数据和趋势，然后通过对每日得分的差分求和，并利用logistic函数来计算产品获得成功的概率。最后我们基于RNTN(递归神经张量网络模型)将评论分成非常消极、消极、中立、积极、非常积极五种类别，使用皮尔逊相关性和斯皮尔曼相关性对评论类别向量和评级向量进行相关性分析。

通过以上分析，提供以下建设性建议：

1、建议商家首先对三种产品的质量进行控制，因为质量在三种产品中的影响占比均比较大。

2、保证价格的适宜度。

3、对产品进行多选择化设计。如微波炉内部空间，整体大小设置可调节适应环境。

4、吹风机的热度，风速调节。

5、吹风机与微波炉电器类的安全要有保障。

6、尽量减少吹风机与微波炉的噪音，但是需要有明显的提示音。

7、外观上设计的更符合大众审美，增加用户体验感。

8、重量体积上需要针对绝大多数人群设计。

参考文献：

[1] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank[C].Proceedings of the 2013 conference on empirical methods in natural language processing,2013: 1631-1642.

[2] 罗胤达.大数据时代下电商自动处理评论文本的研究[J].中国市场，2020 年 36 期：166-167.

[3] Sang-Woon Kim, Joon-Min Gil. Research paper classification systems based on TF-IDF and LDA schemes[J]. Human-centric Computing and Information Sciences, 2019, 9(1):8-9.

[4] Huilong Fan, Yongbin Qin. Research on Text Classification Based on Improved TF-IDF Algorithm[P]. Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018),2018:12-15.

[5] 刘勇,任晓伟.一种基于深度学习的电商平台用户评论情感分类方法[J].青岛科技大学学报(自然科学版),2020 年 06 期：99-106.

[6] 章蓬伟,贾钰峰,邵小青,拜尔娜·木沙,赵裕峰.基于文本情感分析的电商产品评论数据研究[J].微处理机,2020 年 06 期：58-62.

[7] 颜颖.C2C 电子商务物流服务客户满意度评价研究——以淘宝网为例[J].湖南农业大学学报(自然科学版).2013(S1):2-7.

[8] Abdel-Basset Mohamed, Mohamed Mai, Elhoseny Mohamed, Son Le Hoang, Chiclana Francisco, Zaied Abd El-Nasser H. Cosine similarity measures of bipolar neutrosophic set for diagnosis of bipolar disorder diseases.[J]. Artificial intelligence in medicine, 2019,25(2):101.

[9] Cutting Keith. EWMA 2018: a conference to remember.[J].British journal of nursing (Mark Allen Publishing), 2018(Sup12):25-30.

[10] Shangjie Xu,Daniel R Jeske.Weighted EWMA charts for monitoring type I censored Weibull lifetimes[J]. Journal of Quality Technology, 2018, 50(2):36-45.

[11]肖乐,轩辕敏峥,段梦诗.一种基于情感词典与微博文本数据的七情感分类方法[P].中国专利:CN110633367A,2019-12-31:10-19.

[12]刘文远,郭智存,于家新,付闯.基于深度学习的评论文本方面级情感分类方法及系统[P].中国专利：CN202010776165.3,2020-10-30:1-14.

[13]岳丹迪.物流服务评论与顾客购买后行为的相关性分析[J].居舍.2018 年 09 期:179.

[14]彭三春,张云华.基于 RNTN 和 CBOW 的商品评论情感分类[J].计算机工程与设计,2018 年 03 期:861-866.

[15]薛福亮,刘丽芳.基于 TF-IDF 和情感强度的细粒度情感分析——餐饮评论为例[J].信息系统工程,2020 年 03 期: 83-84+86.

[16]李良强,李开明,白梨霏,曹云忠,吴亮.网购农产品评论中的消费者情感标签抽取方法研究[J].电子科技大学学报(社科版).2018(04):25-29.

致谢

感谢我的导师孔老师，不辞辛苦地给予我指导，拓展我的思路，使我的论文能够圆满完成。

感谢我的室友们，是他们给予我成长前进的动力，在遇到棘手问题时不厌其烦的给我帮助，激发我的灵感。

最后感谢我的父母，提供我学习的平台和环境，一直默默的支持着我且毫无怨言，在未来的道路上，我会一直努力下去，将来报答她们的养育之恩。

学不止境，未来的生活会越来越精彩！