分类号 C8/416　　密级 公开
U D C 0005640　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 基于主题建模与情感分析的冰雪大世界在线评论研究

研 究 生 姓 名：徐应发

指导教师姓名、职称：刘 明 教授

学 科、专 业 名 称：应用统计

研 究 方 向：大数据分析

提 交 日 期：2024 年 6 月 3 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：徐方发 签字日期：2024.5.30

导师签名：刘明 签字日期：2024.5.30

导师(校外)签名：_____ 签字日期：_____

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意！（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：徐方发 签字日期：2024.5.30

导师签名：刘明 签字日期：2024.5.30

导师(校外)签名：_____ 签字日期：_____

Research on Online Commentary of the Ice and Snow World Based on Theme Modeling and Emotional Analysis

Candidate : XU Yingfa

Supervisor: Liu Ming

摘 要

随着互联网的普及和社交媒体的兴起，更多的游客选择撰写在线评论，分享自己的旅游体验和感受。这些互联网平台每天都会产生海量的评论数据，挖掘并分析这些评论数据的潜在情感倾向和偏好信息，可以帮助景区管理者了解游客的态度和需求，对于景区的服务改进和体验优化有着重要价值。

后冬奥时代，冰雪运动加速普及，冰雪产业加快发展，冰雪旅游持续升温，兴起了各项冰雪旅游的热潮。北疆“冰城”哈尔滨以其得天独厚的冰雪资源和丰富多彩的冰雪文化而火爆出圈，成为热度最高的话题之一，推动了当地旅游业的蓬勃发展。冰雪大世界作为这座“冰城”的标志性景点，人气火爆，网络评论数据丰富，选择冰雪大世界作为研究对象并收集在线旅游平台上的游客评论数据。

本研究旨在深入分析哈尔滨冰雪大世界的在线评论，首先采用 LDA 模型对评论中的正面与负面情感进行主题分析，总结不同主题词下的主题内容，从而揭示游客的旅游偏好，为景区管理者提供具有针对性的改进策略。然后分别使用情感词典、传统机器学习和深度学习的方法对评论数据进行情感分析，并且为提升情感分析的准确性和丰富性，提出了一种 BERT-EW 双通道情感分析模型。该模型巧妙地将评论的语义特征和情感词特征分开处理，通过语义通道和情感词通道分别提取特征，更有效地捕获评论中更丰富、更精确的情感信息。

根据分析结果，可以得出以下结论：（1）游客的正面评价主要集中在冰雪大世界的观赏性、订票和取票的便利性、门票价格的合理性和优惠政策以及景区交通的便捷性等方面；负面反馈主要针对导游和客服的服务质量、过长的排队等候时间以及门票价格和景区收费高昂等问题。（2）在情感分析中，使用传统机器学习方法要优于情感词典的方法，其中多项式朴素贝叶斯算法的效果较好；使用深度学习的方法进行情感分析效果显著提升，特别是 BERT 模型的效果远高于 Word2Vec 词向量模型。（3）本研究提出的基于 BERT 的双通道情感分析模型 BERT-EW，成功地融合了评论的语义信息和情感词信息，在哈尔滨冰雪大世界评论数据上，该模型的表现效果要优于 BERT-BiGRU 模型。

关键词：冰雪大世界；主题分析；情感分析；LDA 模型；BERT 模型

Abstract

With the popularity of the Internet and the rise of social media, more and more tourists choose to write online reviews to share their travel experiences and feelings. These Internet platforms generate massive comment data every day. Mining and analyzing the potential emotional tendencies and preferences of these comment data can help scenic spot managers understand the attitudes and needs of tourists, which is of great value for the service improvement and experience optimization of scenic spots.

In the post Winter Olympics era, the popularization of ice and snow sports has accelerated, the development of the ice and snow industry has accelerated, and ice and snow tourism has continued to heat up, leading to the rise of various ice and snow tourism trends. Harbin, known as the "Ice City" in northern Xinjiang, has gained popularity due to its unique ice and snow resources and rich and colorful ice and snow culture, becoming one of the hottest topics and promoting the vigorous development of local tourism. As a landmark attraction of this "ice city", the Ice and Snow World is popular and has rich online review data. We chose the Ice and Snow World as the research object and collected tourist review data on online travel platforms.

This study aims to conduct an in-depth analysis of online comments on Harbin Ice and Snow World. Firstly, the LDA model is used to analyze

the positive and negative emotions in the comments, summarize the theme content under different theme words, and reveal tourists' travel preferences, providing targeted improvement strategies for scenic area managers. Then, sentiment analysis was performed on the comment data using sentiment dictionaries, traditional machine learning, and deep learning methods. To improve the accuracy and richness of sentiment analysis, a BERT-EW dual channel sentiment analysis model was proposed. This model cleverly separates the semantic features and emotional word features of comments, extracts features separately through semantic and emotional word channels, and more effectively captures richer and more accurate emotional information in comments.

Based on the analysis results, the following conclusions can be drawn: (1) Positive evaluations from tourists mainly focus on the viewing value of the ice and snow world, the convenience of booking and collecting tickets, the rationality of ticket prices and preferential policies, and the convenience of transportation in scenic areas; Negative feedback mainly targets issues such as the service quality of tour guides and customer service, long waiting times in queues, and high ticket prices and scenic area fees. (2) In sentiment analysis, traditional machine learning methods are superior to sentiment lexicon methods, with polynomial naive Bayes algorithm performing better; The use of deep learning methods for sentiment analysis has significantly improved the performance, especially

the BERT model, which is much better than the Word2Vec word vector model. (3) The dual channel sentiment analysis model BERT-EW based on BERT proposed in this study successfully integrates semantic information and sentiment word information of comments. On the Harbin Ice and Snow World comment data, the performance of this model is superior to the BERT-BiGRU model.

Keywords: Ice and Snow World; Theme Analysis; Emotional Analysis; LDA Model; BERT Model

目 录

1 绪论.....1

1.1 研究背景与意义.....1

1.2 国内外研究现状.....2

1.2.1 旅游在线评论研究现状 2

1.2.2 LDA 主题模型研究现状.....3

1.2.3 情感分析研究现状.....4

1.3 创新点.....6

1.4 研究内容.....7

2 相关理论及基础.....8

2.1 LDA 主题模型.....8

2.2 传统机器学习模型.....9

2.2.1 决策树.....9

2.2.2 逻辑回归.....10

2.2.3 支持向量机.....10

2.2.4 朴素贝叶斯.....11

2.3 神经网络模型.....12

2.3.1 卷积神经网络.....12

2.3.2 长短时记忆神经网络.....13

2.3.3 门控循环单元.....15

3 冰雪大世界在线评论主题分析.....16

3.1 哈尔滨冰雪大世界现状.....16

3.2 数据获取与预处理.....17

3.2.1 数据来源.....17

3.2.2 评论数据采集.....17

3.2.3 评论数据预处理.....19

3.3 在线评论基本分析.....21

3.3.1 描述性统计分析.....21
3.3.2 词频分析.....23
3.4 LDA 主题分析.....25
3.4.1 LDA 主题建模.....26
3.4.2 主题个数分析.....27
3.4.3 主题结果分析.....29
4 冰雪大世界在线评论情感分析.....33
4.1 基于情感词典的情感分析.....33
4.1.1 情感词典的构建.....33
4.1.2 评分规则.....36
4.1.3 实验数据与评价指标.....36
4.1.4 结果分析.....38
4.2 基于传统机器学习的情感分析.....38
4.2.1 特征提取.....39
4.2.2 模型训练.....39
4.2.3 结果分析.....41
4.3 基于深度学习的情感分析.....42
4.3.1 词嵌入模型.....42
4.3.2 BERT 预训练模型.....43
4.3.3 模型训练.....45
4.3.4 结果分析.....46
4.4 基于 BERT-EW 的双通道情感分析模型.....47
4.4.1 BERT-EW 双通道模型.....47
4.4.2 参数设置.....50
4.4.3 结果分析.....50
5 结论与展望.....52
5.1 结论.....52
5.2 建议.....53
5.3 不足与展望.....53

参考文献……55

致谢……59

1 绪 论

1.1 研究背景与意义

1996 年建立的原旅游局信息中心标志着中国旅游业信息化的起步，并逐步推动旅游业的数字化进程。2006 年，科技部启动并实施的“数字旅游项目计划”进一步推动了旅游业的数字化发展，这个计划旨在利用数字技术提升旅游资源的开发和管理，改善游客体验，并推动旅游业的信息共享和服务创新。到了 2015 年，国家旅游局开启了“旅游+互联网”的项目计划，进一步深化互联网与旅游的联系合作，通过技术革新改变旅游业的生产方式和发展模式，优化网络营销，以及提高旅游公共服务的效率和质量。这些发展反映了中国旅游业在面向数字化转型的过程中，不断采纳新技术和互联网应用，以适应消费者行为的变化和市场需求，数字技术的应用使得旅游服务更加个性化、便捷，并提供了更为丰富的旅游产品和服务。

随着数字旅游时代的到来，加速了互联网和旅游行业的深入融合，信息技术与旅游资源紧密结合，并广泛应用于各个领域。在这个环境下，游客出行借助互联网进行旅游成为常态，旅游信息的获取和分享变得前所未有的便捷，这使得在线评论成为潜在游客获取旅游信息和制定决策的重要渠道。自从八大 OTA 平台在 2018 年年底召开的线上旅游数据生态及治理高峰会上发布了《在线旅游内容和数据生态共建》倡议，网上评论数据能够更真实、准确地反映旅客的真实情感，成为企业和消费者获取信息的重要渠道，帮助商家根据旅客需求提高服务水平以增强竞争力。

旅游业是全球最大的行业之一，对许多国家和地区的经济具有显著影响，特别是冰雪旅游业，以其独特的冰雪资源，成功吸引了大量热衷于冰雪运动与冰雪文化体验的游客，为我国冰雪旅游产业带来了新的发展机遇。2021 年 2 月，文化和旅游部、国家发展改革委、国家体育总局联合印发《冰雪旅游发展行动计划（2021—2023 年）》。3 年来，全国各地深入践行“冰天雪地也是金山银山”理念，推动冰雪旅游与相关行业融合，有效促进冰雪旅游业的发展，提升整个产业链的竞争力和效益，切实推动冰雪旅游高质量发展。2022 年北京冬奥会的成功举办无疑是一个重要的里程碑，成功点燃了大众参与冰雪运动的热情，借筹办此次冬

奥会契机，我国不仅顺利实现了“三亿人参与冰雪运动”的目标，也吸引了更多游客和投资者的关注，为冰雪旅游业的发展注入了新的活力和动力。中国旅游研究院发布的《中国冰雪旅游发展报告（2023）》显示，冰雪旅游在过去几年取得了显著的增长，从 2021 年到 2022 年的冰雪季，全国冰雪休闲旅游人数达到了 3.44 亿人次，是 2016 年到 2017 年冰雪季人次的两倍多，冰雪休闲旅游收入达到了 4740 亿元，2022 年到 2023 年的冰雪季全国冰雪休闲旅游人数达 3.12 亿人次，中国正从冰雪旅游体验阶段进入冰雪旅游刚性生活需求阶段。

纵观全国，今冬最为火爆的冰雪景区非哈尔滨冰雪大世界莫属，冰雪大世界作为哈尔滨的一张城市名片，因其独特的冰雪景观和丰富的冰雪活动而备受游客喜爱，已成为中外游客冬季旅游的热门打卡地。冰雪大世界每天都吸引着数以万计来自各地的游客，通过这些游客发表的在线评论能够帮助景区更好地了解游客需求，提高游客满意度，对景区管理部门做出改善决策起到了重要作用。

对冰雪大世界旅游景点在线评论的研究旨在通过分析游客的评论，了解他们对景点的评价和情感倾向，以便更迅速地洞悉游客的需求和喜好，做出更明智的决策。主题建模和情感分析是文本挖掘中常用的技术工具，可以帮助研究者从大量的评论文本中提取出关键主题并理解游客的情感倾向。主题建模能识别评论中的重要话题和关键词，帮助了解游客对于景区的关注焦点和评价侧重点，为景区的产品开发和服务优化提供参考和方向。情感分析则有助于判断评论的情感极性，即评价者对景区的态度是正面、负面还是中立，从而使景区管理者更好地了解游客的满意程度和不满之处，以便及时调整服务并改进游客体验。通过主题建模和情感分析，能够更深入地理解游客对哈尔滨冰雪大世界景点的评价和反馈，为景区提供更有针对性的改进建议和服务优化方案。

1.2 国内外研究现状

本文的研究问题是对游客的在线评论进行挖掘分析，因此，国内外研究现状从旅游在线评论研究、LDA（Latent Dirichlet Allocation）主题模型以及情感分析这三个方面进行阐述。

1.2.1 旅游在线评论研究现状

旅游在线评论是旅游业市场营销和客户体验管理的重要组成部分，它不仅影响潜在旅客的旅行决策，还为旅游服务提供商提供宝贵的反馈，国内外研究者对

此领域进行了广泛研究。旅游在线评论对消费者的旅游决策有显著影响，包括选择目的地、酒店、景点和旅行社等，研究者探讨了评论的数量、质量、情感倾向和来源等因素如何影响消费者的决策。Torres 等（2015）探讨了旅游在线评论对酒店市场份额造成的变化，发现旅游在线评论的评分和评论数量对酒店在线预订收入产生显著影响。陆之洲（2021）的细粒度情感分析揭示了旅游在线评论情感与销量之间的相关性，强调了负面评论的较大影响力，并指出细粒度情感提供了比粗粒度更准确的解释力。

文本挖掘和情感分析是对旅游在线评论的主要研究方法，文本挖掘用于提取评论中的关键信息，如主题、观点和关键词等。Fazzolari 等（2018）收集了意大利某个景区跨越 8 年的在线评论，指出相关的旅游平台可以通过在线评论获取有价值的信息、制作针对性较强的行动策略。王承云等（2022）基于携程网、马蜂窝网两大 OTA 中上海红色旅游目的地的在线评论，使用词云词频统计法、复杂网络分析法及扎根理论分析法，对上海红色旅游形象感知与情感评价进行研究。司育（2023）利用 LDA 模型挖掘在线评论的潜在主题，并根据各个主题下的关键词对主题进行提炼命名，得到相对客观的旅游目的地评价指标，并对各指标的主题得分进行测算，从而实现山西省十大热门景区的综合性评价。

旅游在线评论的情感分析则用于判断评论的情感倾向，即评论者对旅游产品或服务的态度是正面还是负面的。涂海丽等（2016）利用领域本体构建方法构建旅游本体，将处理后的评论文本与旅游本体进行匹配，得出本体各属性的分类评论集，运用情感程度加权规则计算这些评论集的情感极性均值，得出游客关于旅游各要素总体情感倾向，并进行可视化分析与展示。刘逸等（2017）借助网络大数据研究分析方法，基于游客情感分析理论，以赴澳中国游客发布在国内旅游网站的评论为素材展开分析，比较其与国际游客的差异性，继而解析主要影响因素，开拓了研究游客偏好和评价的新方法。严仲培等（2019）基于词向量模型，提出一种情感词典种子词集筛选方法，将情感词语以向量形式表征并计算词向量间距离，形成种子词集的筛选标准和分类依据，再通过类别判断形成了山岳型旅游景区在线评论情感词典。

1.2.2 LDA 主题模型研究现状

LDA 主题模型是学者 Blei 等（2003）提出的一种文本分析模型，该模型的

提出有效弥补了概率隐形语义分析模型（Probabilistic Latent Semantic Analysis，PLSA）的缺陷，提升了对深层次文本内容挖掘和语义分析的精准度与效率。尽管基础的 LDA 模型已经非常强大，但研究者们还提出了许多扩展和改进模型，以解决特定的问题。在线文本具有明显的时间属性，针对动态文本建模问题，Alsuraait 等（2008）提游一种在线 LDA 模型（On-Line LDA），当有新的文本流更新的时候，该模型可以利用已得出的主题模型增量式地更新当前模型，不再需要重新访问之前所有的数据，能够实时获取随时间变化的生题结构。张晨逸等（2011）提出了一个基于 LDA 的微博生成模型 MB-LDA，综合考虑了微博的联系人关联关系和文本关联关系，来辅助进行微博的主题挖掘。采用吉布斯抽样法对模型进行推导，不仅能挖掘出微博的主题，还能挖掘出联系人关注的主题。Das 等（2015）最早尝试从词向量空间中采样主题，提出 GLDA（Gaussian LDA）模型，在该模型中，假设文档不是由单词类型序列组成的，而是由单词嵌人向量组成的。基于词向量的概率主题模型均直接利用事先训练的词向量来辅助模型的学习，使得语义相近的词汇依较大概率获得同一主题，提高主题词的一致性和可解释性，丰富了文本的潜在特征表达，进而有效地提高模型分类的准确性。刘干等（2021）通过引入中心词概念，提出一种改进 LDA 主题模型，通过对比传统 LDA 模型和改进 LDA 模型，发现改进方法所生成的 LDA 模型在高频词分布集中度上更优于传统方法，在下游任务应用中更适合热点话题生成。王晨等（2023）首先提出新的潜在特征主题模型 Improved-LDA，以个人隐私信息法律保护领域的研究文献为例，在主题一致性计算与主题聚类方面，Improved-LDA 模型的性能明显优于传统的 LDA 模型。阮光册等（2023）将 Sentence-BERT 句子嵌入模型和 LDA 模型相结合，提升评论文本主题的语义性，增加了模型的复杂性，该方法获得的主题一致性指标优于目前常见的评论文本主题识别方法。

1.2.3 情感分析研究现状

早期，大多数情感分析基于语义规则的情感词典或传统机器学习方式，通过计算情感词的权重得出情感倾向。肖红等（2014）提出了一种将句法分析和情感词典相结合的分析方法，在用情感词典进行切词的基础上对句子进行语法分析，再利用情感词在句子中的成份、情感指数权重以及与其他情感词之间的组合共现关系计算出综合的情感指数。为提高情感分类的准确性，研究者对基于传统机器

学习的方法进行了研究，取得了不错的结果。Pang 等（2002）是首个将机器学习应用在情感分析任务中的，使用 SVM（Support Vector Machine）、朴素叶斯和最大熵等算法来分析电影评论情感倾向，实验表明文本特征与 SVM 算法组合的效果更佳。唐慧丰等（2007）通过使用几种常见的机器学习方法（SVM、KNN 等）对中文文本的情感分类进行了实验比较，通过大量的对比实验发现采用 BiGrams 特征表示方法、信息增益特征选择方法和 SVM 方法时，在大量训练集和适量的特征选择时情感分类效果达到最优。

随着词向量模型的提出与广泛应用，基于词向量的深度学习语言模型开始展现出了强大的分类能力，基于此，国内外的科研学者们开始将深度神经网络算法引入到情感分类模型构建中以提升模型性能。性能表现较好的情感分类模型大多以循环神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Network，CNN）为基础构建。如刘龙飞等（2015）分别将字级别词向量和词级别词向量作为原始特征，采用卷积神经网络提取文本特征，在任务语料上进行了情感分析实验，取得良好效果。相比之下，循环神经网络在处理文本数据时具有优势，因为它能够捕捉序列数据中的长距离依赖关系，但当输入数据中存在长期依赖关系时，它们会出现梯度消失和梯度爆炸，而长短时记忆网络（Long Short-Term Memory，LSTM）可以很好的解决长期依赖的关系。Tang 等（2015）以长短时记忆网络为基础，使用微博评论文本作为训练集，训练得到短文本情感分类模型，该情感分类模型可对未知短文本语料进行情感倾向的判别。张仰森等（2018）采用双向长短记忆网络模型和全连接网络，分别对微博文本和文本中包含的情感符号进行编码，有效增强了情感语义捕获能力，提高了微博情感分类的性能。这些基于 RNN 变体的 LSTM 模型能够捕捉到较长距离的依赖关系，但仍然存在无法编码从后到前的问题。针对该问题，Zhang 等（2018）提出一种基于双向递归神经网络的分层多输入输出模型，该模型采用两个 GRU（Gated Recurrent Unit）来获取词性和句子的词向量表示，加快了多标签情感识别的计算效率。深度神经网络在情感分析上取得了显著的成果，但其考虑的是句子中所有的词，不能关注文本的突出部分。通过在深度学习的方法中加入注意力机制，用于情感分析任务的研究，能够更好地捕获上下文相关信息，有效提高文本情感分类的准确率。胡荣磊等（2019）使用 Word2Vec 进行词向量表示，输入长短时记

忆网络，再通过注意力机制分配权重来进行文本情感分析。李磊等（2021）将对象信息与文本信息进行融合，利用注意力机制强化的 BiLSTM 模型得到评论文本的情感分类结果。

上述基于词向量技术的神经网络模型在情感分析任务中取得了不错的分类效果，但 Word2Vec 和 GloVe（Global Vectors）等常用词向量技术，其生成的词向量是静态的，集中于获得词语浅层特征表示，无法进一步解决相同词语在不同的场景下多义性的问题。Devlin 等（2018）提出 BERT（Bidirectional Encoder Representations from Transformers）预训练模型，使用深度双向 Transformer 模型，随时根据上下文信息动态调整词向量，解决了传统语言模型存在的一词多义问题，在提升模型性能的同时也大幅度降低了训练难度以及训练所耗费的时间。游兰等（2023）针对传统情感识别模型大多集中于评论的表层语义挖掘，存在分类效果不佳、泛化能力有限等问题，提出了一种基于 BERT-BiGRU 多模集成学习的深层情感语义识别方法，相较于其他传统模型有更优的情感识别效果。诸林云等（2023）利用 BERT 预训练语言模型获取用户对酒店评论的文本特征表示，使用双向 LSTM 考虑文本中过去和未来的上下文依赖关系，并对文本的不同部分给予不同的关注，从而提高情感分类的准确度。

综上所述，在数字化和信息技术飞速发展的背景下，情感分析和主题建模等技术在旅游行业中的应用日益受到全球学者的关注。中国的冰雪旅游业正在经历一个快速增长的阶段，特别是在冬季奥运会带动下，冰雪运动和旅游的普及和发展吸引了国内外游客的极大兴趣。然而，相较于其他领域，冰雪旅游领域的在线评论研究还相对较少，因此本文将对哈尔滨冰雪大世界的在线评论进行主题建模，以揭示游客评论中的主要话题和趋势，并通过将情感词典与深度学习技术结合进行情感分析，更准确地捕捉和理解游客的情感倾向。

1.3 创新点

在研究方法上，为提升情感分析的准确性和丰富性，结合情感词典与深度学习方法提出了一种 BERT-EW（BERT-Emotional Words）双通道情感分析模型，该模型通过构建情感词典提取评论文本的情感词特征，巧妙地将评论的语义特征和情感词特征分开处理，使用 BERT-BiGRU 模型分别对语义通道和情感词通道进行特征提取，更有效地捕获评论中更丰富、更精确的情感信息。

1.4 研究内容

本文以旅游网站上哈尔滨冰雪大世界这一著名冰雪景点的游客在线评论数据作为研究的数据来源，爬取携程、同城和去哪儿三个旅游网站上的评论数据，将获得的文本数据进行清洗、文本分词，然后进行描述性统计分析和主题分析，识别出评论中涉及的主题和话题，最后对评论数据进行情感分析。本论文分为五章，主要研究内容和章节安排如下：

第一章，阐述本文的研究背景和意义，然后以旅游在线评论研究内容和使用的分析方法不同，分别从旅游在线评论分析、LDA 主题模型和情感分析三个方面，展开介绍了游客在线评论分析的研究进展与研究现状，最后给出了本文的主要研究工作和各章节安排。

第二章，重点对文中所涉及的基础理论进行了介绍，首先介绍 LDA 主题模型及原理，然后介绍了传统机器学习模型，最后对构建情感分析模型中常用的深度学习相关技术进行了阐述，其中介绍了卷积神经网络、循环神经网络，为下文的研究打下了坚实的理论基础。

第三章，对哈尔滨冰雪大世界进行了简单介绍，然后使用 python 爬虫采集冰雪大世界在线评论，对获取的数据进行预处理，接着对评论数据进行文本分析，包括描述性统计分析和词频分析，最后使用 LDA 主题模型对评论数据进行主题分析，提取评论的主题特征。

第四章，分别使用情感词典、传统机器学习和深度学习的方法对评论数据进行情感分析，然后在 BERT 模型基础上融合情感词特征，提出了 BERT-EW 情感分析模型，并对模型的各组成部分进行了简要描述，最后搭建了 BERT-EW 模型并验证其有效性。

第五章，对本文的研究工作进行了总结，然后针对研究成果给冰雪大世界景区提出一些建议，最后分析了存在的不足与改进的方向。

2 相关理论及基础

本章主要介绍本文所使用到的文本分析相关技术，包括 LDA 主题建模、传统机器学习模型和神经网络模型，为后续的实证研究奠定理论基础。

2.1 LDA 主题模型

LDA 主题模型是一种用于文本挖掘和自然语言处理的概率生成模型，由 Blei，DavidM，Ng，AndrewY，Jordan 于 2003 年提出。LDA 主题模型能够从大量的非结构化文本数据中推测出文档的主题分布，将一篇文档的主题分布通过概率分布的形式输出，从而了解文档中涉及的主题内容，并且通过推断出文档的主题分布，可以实现主题聚合或文本分类，为文本数据的分析和理解提供更多的可能性。

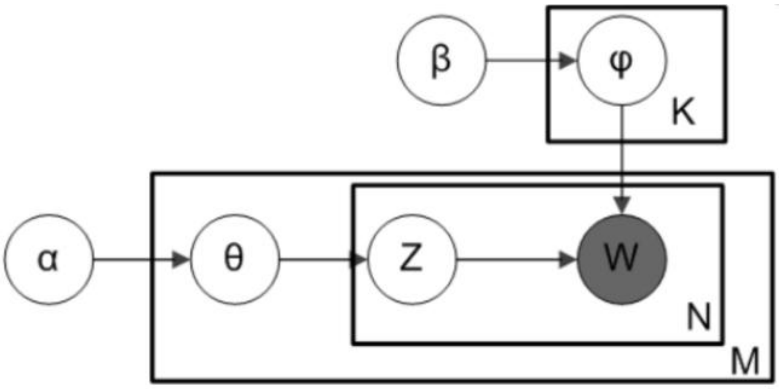

图 2.1 LDA 主题模型概率图

在 LDA 主题模型中，每个文档的主题分布 P(z|d)和每个主题中单词出现的概率 P(w|z)被视为按照迪利克雷分布（Dirichlet Distribution）概率随机的变量。图 2.1 是 LDA 主题模型的概率图，生成一篇文档的步骤如下：

（1）对狄利克雷分布α进行采样，随机生成文档对应主题的多项式分布θ；

（2）对主题的多项式分布θ进行采样，随机生成一个主题 z；

（3）对狄利克雷分布β进行采样，随机生成主题对应词语多项式分布φ；

（4）综合主题 z 和主题对应词语分布情况φ生成词语 w；

（5）不断循环第四步生成一个文档，包涵 n 个词语，最终生成 k 个主题下的 m 篇文档。

在使用 LDA 主题模型进行主题分析时，主题的个数本质上等价于文本聚类中的聚类个数，确定主题个数 K 是一个重要的步骤，因为选择不合适的主题个数可能会影响模型的性能和结果的解释性。一般来说，确定主题个数 K 的方法有两种方式，计算困惑度（perplexity）或一致性指标（coherence）。

2.2 传统机器学习模型

2.2.1 决策树

决策树模型是一种监督学习算法，主要用于分类和回归问题，常见的决策树算法有 ID3、C4.5、CART（Classification and Regression Trees）等。它以树状图的形式来表示决策规则和可能的结果，其中树中的每一个内部节点代表一个特征属性上的测试，每个分支代表测试的结果，而每个叶节点代表一种类别或者一个输出值，如图 2.2 所示。

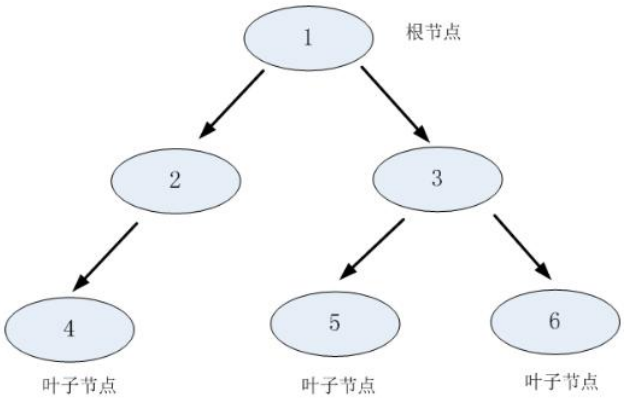

图 2.2 决策树模型图

决策树模型的构建过程通常包括以下步骤：（1）选择一个最佳特征进行分割，最佳特征的选择可以采用信息增益、增益率、基尼指数等方法；（2）根据选择的特征，将数据集划分为不同的子集，递归地生成子树；（3）对决策树进行枝剪优化，以防止过拟合，剪枝方法包括预剪枝和后剪枝，预剪枝是在构建过程中提前停止树的生长，后剪枝是在构建完整的决策树后进行简化。

决策树模型的主要优点是其可解释性强，因为它生成的决策规则非常直观，易于理解。其他的机器学习算法需要对数据进行复杂的预处理，如归一化、缺失

值处理等，而决策树模型则不需要这些复杂的预处理步骤，并且决策树模型可以处理多输出问题，即一个实例可能同时属于多个类别。

虽然决策树模型有诸多优点，但是它也存在一些局限性。决策树模型容易过拟合，特别是在处理包含大量特征和复杂结构的数据时，此外，决策树模型对于连续特征的处理不是很理想，因为它采用的是二分的方式来处理连续特征，这可能会导致模型性能下降。

2.2.2 逻辑回归

逻辑回归模型是一种广泛应用于统计学和机器学习领域的分类算法，尤其适用于二分类问题，尽管其名称含有“回归”，逻辑回归实际上是一种分类技术，用于预测一个事件的发生概率。它通过使用逻辑函数将线性回归的输出映射到 0 和 1 之间，因此输出可以被解释为概率，逻辑函数通常使用 Sigmoid 函数，Sigmoid 函数的公式为：

$$\text{sigmoid}(x)=\frac{1}{1+e^{-x}} \tag{2-1}$$

逻辑回归模型通过设定一个阈值，将 Sigmoid 函数的输出转换为二元分类结果，如果某个样本的预测概率大于或等于这个阈值，则将其分类为正类，否则为负类。

逻辑回归模型易于实现和理解，适用于解释变量与二元结果之间的关系，模型的输出能通过概率直观地解释，每个特征的权重表示了该特征对最终概率的贡献大小和方向。虽然逻辑回归可以通过一对多策略扩展到多类分类，但逻辑回归天然适用于二分类问题。

2.2.3 支持向量机

支持向量机（SVM）是一种强大的监督式学习模型，广泛应用于分类、回归及异常检测任务。SVM 特别适合应用于中小规模数据集的复杂模式识别中，能够处理高维特征空间的数据，甚至在特征数量超过样本数量的情况下也表现出色。

SVM 的基本想法是找到一个超平面来分隔不同类别的数据，同时使得最近的数据点到这个超平面的距离尽可能大，在最简单的二分类问题中，这个超平面可以直观理解为一条直线或一个平面，其目的是创建一个最佳的决策边界，如图

2.3 所示。

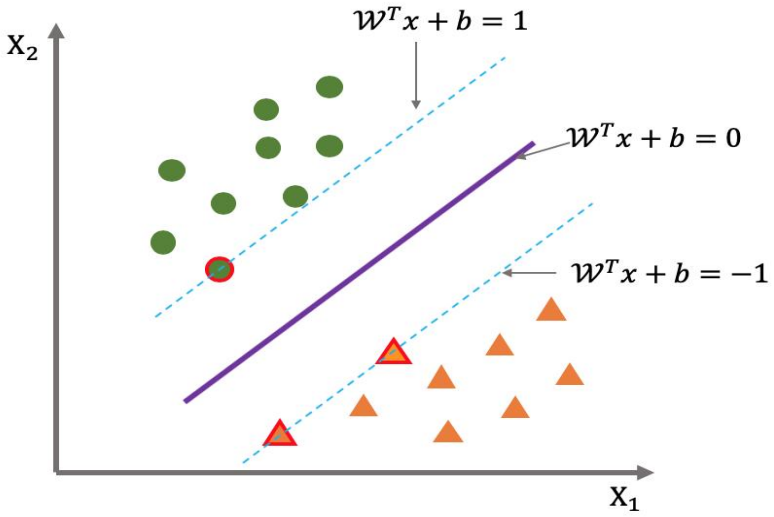

图 2.3 支持向量机模型图

决策边界的确定完全依赖于距离它最近的那些数据点，这些点被称为支持向量，支持向量是决策边界最终位置的关键，任何不是支持向量的数据点都可以不考虑，它们对分类边界没有影响。在实际应用中，很多数据集不是线性可分的，SVM 通过引入核技巧将数据映射到一个更高维的空间，使得数据在这个新空间中线性可分。

SVM 在多种数据集上表现出良好的泛化能力，即使在特征数远大于样本数的情况下，SVM 也能有效工作。但 SVM 的性能高度依赖于核函数的选择及其参数和正则化参数 C 的设定，对于非常大的数据集，SVM 的训练时间可能很长，计算成本高。

2.2.4 朴素贝叶斯

朴素贝叶斯（Naive Bayes）模型是一种基于贝叶斯定理的分类方法，以其简单性和效率在机器学习领域内被广泛应用。该模型的“朴素”二字来源于它对特征间相互独立的假设，即假设各特征之间不存在任何依赖关系。

朴素贝叶斯模型基于贝叶斯定理计算给定数据样本属于某个类别的后验概率，主要包括计算各类别的先验概率以及给定类别下各特征的条件概率，通过对训练数据进行简单的频率统计即可获得这些概率。

贝叶斯定理公式如(2-2)所示：

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)} \tag{2-2}$$

其中，A_i是目标类别，B 是给定的数据样本，$P(A_i \mid B)$ 是在已知数据样本 B 的条件下A_i类别的后验概率，$P(B \mid A_j)$是给定类别A_i下数据样本B的概率。

朴素贝叶斯模型有三种主要的变体，包括高斯朴素贝叶斯（Gaussian Naive Bayes）、多项式朴素贝叶斯（Multinomial Naive Bayes）以及伯努利朴素贝叶斯（Bernoulli Naive Bayes）。高斯朴素贝叶斯假设每个类别的连续特征都符合高斯分布，适合处理连续数据，多项式朴素贝叶斯和伯努利朴素贝叶斯则更适合处理离散数据。多项式朴素贝叶斯的计算粒度是词语，它基于每个类别中各个词语出现的频率来计算概率，特别适合处理文本数据，常被应用于文本分类任务。伯努利朴素贝叶斯则是在文档级别上进行计算，它关注的是每个词语是否在文档中出现，而不是出现的次数。

2.3 神经网络模型

深度学习的核心思想是通过多层神经网络来学习数据的特征表示，每一层神经网络都由多个神经元组成，每个神经元都接收上一层神经元的输出，并通过激活函数进行非线性变换，通过多层神经网络的组合，深度学习可以学习到更加复杂和抽象的特征表示。本节主要介绍其中的卷积神经网络和循环神经网络这两类神经网络模型。

2.3.1 卷积神经网络

卷积神经网络（CNN）是一种专门用于处理具有网格结构数据的神经网络模型，它在计算机视觉领域中被广泛应用于图像识别、目标检测、图像分割等任务。卷积神经网络的核心是卷积层，通过卷积操作对输入数据进行特征提取，卷积操作可以看作是一种滑动窗口的方式，通过在输入数据上滑动卷积核（也称为滤波器）来提取局部特征。卷积神经网络的优势在于它能够自动学习图像中的局部特征，并且具有平移不变性，这意味着即使目标在图像中的位置发生变化，卷积神经网络仍然能够正确识别。

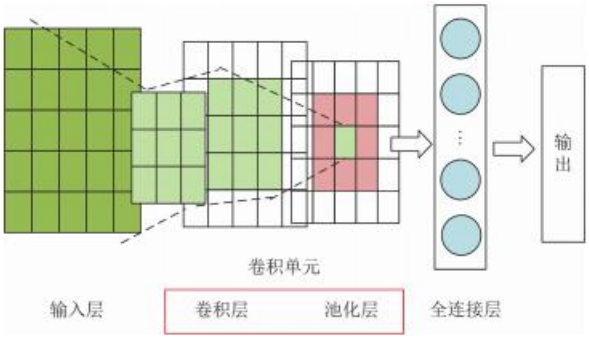

图 2.4 CNN 模型结构

图 2.4 为卷积神经网络示意图，卷积层通过卷积核提取输入图像的不同特征，池化层位于卷积层之后，通过池化操作减少卷积层间的连接数量，进行降维和二次特征提取，两者合称为卷积单元，最后经过全连接层汇总所有卷积单元提取的局部特征，将特征图映射到最终的输出类别。

卷积层、池化层的计算过程分别如式(2-3)、(2-4)所示：

$$C = f(X \otimes W + b) \tag{2-3}$$

$$Z = \text{subsampling(C)} \tag{2-4}$$

其中 C 表示卷积层，W 表示权值向量，$\otimes$ 表示卷积运算，b 表示偏置参数，$f()$ 表示激励函数，Z 表示池化层，subsampling 表示采样过程。

2.3.2 长短时记忆神经网络

长短时记忆神经网络（LSTM）是一种特殊的循环神经网络，设计目的是解决传统 RNN 在处理长序列时容易出现梯度消失或梯度爆炸的问题。LSTM 引入了一个称为“记忆单元”的结构，它可以在不同时间步之间传递和保存信息，记忆单元由一个遗忘门、输入门和输出门组成，通过控制这些门的开关状态，LSTM 可以选择性地忘记、存储和读取信息，这种门控机制使得 LSTM 能够有效地处理长期依赖关系，从而更好地捕捉序列数据中的重要特征。LSTM 的结构如图 2.5 所示：

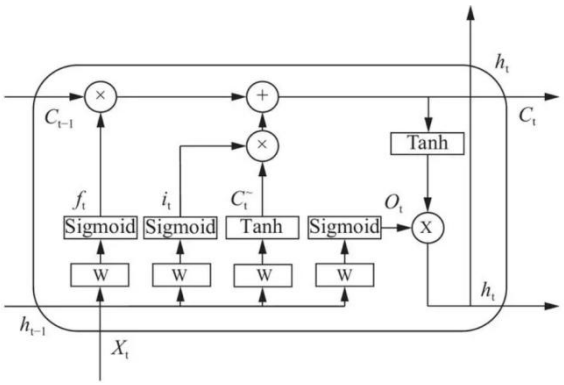

图 2.5 LSTM 模型结构

遗忘门，决定了上一个时间步的记忆单元中哪些信息需要被遗忘，计算过程如式(2-5)所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2-5}$$

输入门，决定了当前时间步的输入信息中哪些需要被存储，计算过程如式(2-6)、(2-7)所示：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2-6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{2-7}$$

细胞状态更新，计算过程如式(2-8)所示：

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{2-8}$$

输出门，决定了当前时间步的输出信息中哪些需要被传递给下一个时间步，计算过程如式(2-9)、(2-10)所示：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2-9}$$

$$h_t = o_t * \tanh(C_t) \tag{2-10}$$

其中 x_t 是神经元的输入，h_t 是神经元的输出，i_t、f_t 和 o_t 是神经元中门的三种状态，C_t 是神经元的输入状态，W 和 b 分别为神经元的输入权重和偏置，σ 代表 sigmoid 函数，tanh 表示 tanh 函数。

2.3.3 门控循环单元

门控循环单元（GRU）是 RNN 的一个变体，能够有效缓解长期记忆和反向传播中的梯度爆炸或弥散等问题，它在保持了 LSTM 的效果的同时又使结构更加简单，是一种非常热门的循环神经网络。GRU 模型由更新门和重置门构成，通过控制这些门的开关状态，GRU 可以选择性地更新和重置隐藏状态。更新门决定了当前时间步的输入信息和前一个时间步的隐藏状态之间的权重，从而控制了隐藏状态的更新程度，重置门决定了当前时间步的输入信息和前一个时间步的隐藏状态之间的权重，从而控制了隐藏状态的重置程度，通过引入更新门和重置门的机制，GRU 能够更好地捕捉序列数据中的重要特征。模型结构如图 2.6 所示：

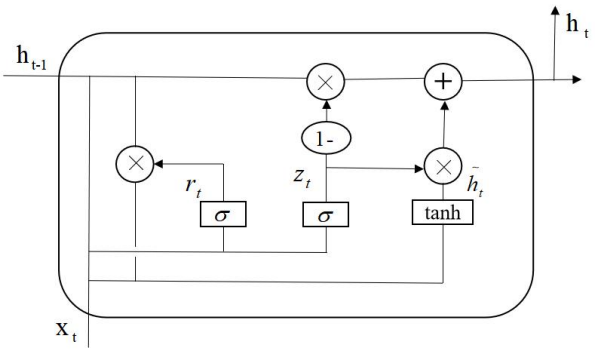

图 2.6 GRU 模型结构

GRU 的具体计算过程如式(2-11)-(2-14)所示：

$$r_t = \sigma(U_r x_t + W_r h_{t-1} + b_r) \tag{2-11}$$

$$z_t = \sigma(U_z x_t + W_z h_{t-1} + b_z) \tag{2-12}$$

$$\tilde{h}_t = \tanh(U_h x_t + r_t * (W_h h_{t-1}) + b_h) \tag{2-13}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{2-14}$$

其中 x_t 为 t 时刻的输入特征，h_{t-1} 为上一时刻的隐藏层状态，U 、W 和 b 分别为权重和偏置，σ 为 Sigmoid 激活函数，*表示矩阵对应元素相乘。

3 冰雪大世界在线评论主题分析

冰雪大世界是中国最大的冰雪主题游乐园，它每年冬季都会举办海内外著名的冰雪艺术展览和冰雪嘉年华活动，吸引了无数游客前来观赏。通过对冰雪大世界的在线评论进行分析，帮助了解游客对景点的需求和关注点，并且使用 LDA 主题模型有效地提取评论中的主题特征，发现不同主题和话题，更加全面地了解游客的评价和反馈。

3.1 哈尔滨冰雪大世界现状

哈尔滨冰雪大世界是位于中国黑龙江省哈尔滨市的一处大型冰雪主题公园，它始创于 1999 年，是为了迎接千年庆典神州世纪游活动而推出的一个大型冰雪艺术精品工程。哈尔滨冰雪大世界凭借哈尔滨的冰雪资源优势，利用了哈尔滨当地在冬季冰雪丰富的特点，每年冬季都会吸引数百万国内外游客前来参观游览。

哈尔滨冰雪大世界的占地面积约为 465 万平方米，主要由数千吨的冰雪构建而成，这里拥有丰富多彩的冰雪雕塑、冰屋、滑梯、冰迷宫以及冰上游乐设施，形成了一个集观赏、娱乐和体验于一体的综合性冰雪旅游景区。景区内包含多个主题区域，如城堡广场、北极熊王国、海底世界、世界之巅、长城等地标性建筑，每个区域都有其独特风格的冰雪雕塑和主题活动。此外，哈尔滨冰雪大世界还提供了一系列的夜间活动和文化表演，以增加游客的参与度和节日气氛。哈尔滨冰雪大世界是一个集冰雪文化、冰雪艺术和冰雪娱乐于一体的景点，是哈尔滨冬季旅游的必去之地，也是中国冰雪文化的重要代表之一。

第二十五届哈尔滨冰雪大世界于 2023 年 12 月 18 日开园迎客，以“龙腾冰雪逐梦亚冬”为主题，将 2025 年哈尔滨亚冬会与龙江地域文化、冰雪文化结合起来进行创作，为世界各地游客打造一座集冰雪艺术、冰雪文化、冰雪演艺、冰雪建筑、冰雪活动、冰雪体育于一体的冰雪乐园。第二十五届哈尔滨冰雪大世界迎来历史最大规模，开园不到 3 小时，预约游玩人数已达 4 万，截止 2024 年 2 月 15 日 24 时正式闭园，共计营业 61 天，累计接待游客 271 万人次。向海内外游客展示了中国东北地区的冰雪魅力，证明了哈尔滨冰雪大世界在我国冰雪旅游领域的领先地位和无与伦比的魅力，刷新了人们对冰雪旅游的认知。

3.2 数据获取与预处理

3.2.1 数据来源

随着互联网技术的日益成熟，在线旅游行业快速发展，人们在选择旅行目的地、酒店、景点等时会参考他人的旅游评论和评价。通过阅读其他游客的经验和建议，能够更好地了解目的地的实际情况，避免一些可能的问题，也可以发现一些隐藏的宝藏景点或特色酒店。

旅游网是一个在线平台，为用户提供包括机票预订、酒店预订、度假产品、火车票、汽车票、旅游度假、门票等在内的全方位旅行服务。旅游网站或平台上还有游客对目的地、酒店、景点等旅游相关内容的评价和评论，这些评价通常包括游客的真实体验、感受和建议，可以帮助其他游客更好地了解和选择旅行目的地、住宿和活动，做出更明智的旅行决策，提高旅行的满意度和体验质量。

目前，国内在线旅游平台主要有携程旅游、去哪儿网、同程旅行、途牛、飞猪旅行等。2024 年 MAIGOO 品牌榜 发布了新的旅游网品牌榜，该榜单由 CNPP 大数据平台提供数据支持，综合分析了旅游网行业品牌的知名度、员工数量、企业资产规模与经营情况等各项实力数据。通过榜单得知排名前十分别是携程、同城旅行、飞猪旅行、去哪儿、马蜂窝、途牛、艺龙旅行、穷游、客路、缤客，其中携程、同程旅行、飞猪旅行和去哪儿分别位列投票榜前列，受到众多网友的喜爱。鉴于飞猪旅行平台在出行和酒店预订方面的数据丰富，但在旅游景点的评论数据较为匮乏，本文决定选取携程、去哪儿和同程旅行三个主要的在线旅游服务平台，作为研究数据的来源。

3.2.2 评论数据采集

为了节省获取游客在线评论数据的时间和成本，使用爬虫自动化地从多个旅游网站上抓取评论数据。相较于依靠人工采集，爬虫可以快速地从大量网页上抓取评论数据，实现大规模数据采集，并且爬虫可以减少人为因素对数据的干扰，提高数据的准确性和可靠性。网络爬虫常用到的软件有 Python、C++或者 Java 等，但是相较于其他的软件，Python 更加简单快捷，本文将采取 Python 爬虫的方式从携程、去哪儿和同程旅行网站获取哈尔滨冰雪大世界景点的在线评论文本数据，包括游客点评时间、IP 属地、评论内容、评分，采集的各个网站的评论数

量如下表 3.1 所示。

表 3.1 评论爬取数量

旅游网站	评论数量
携程	4502
去哪儿	795
同城	8089
合计	13386

本次共采集到哈尔滨冰雪大世界在线评论数据 13386 条，数据内容包括游客点评时间、IP 属地、评论内容、评论分数，爬取的部分数据内容如表 3.2 所示。截止到 2024-02-22，携程网冰雪大世界景点共有 28089 条评论，因为只显示了 300 页的点评信息，所以通过筛选排序爬取了前 300 页的在线评论，共 4502 条。去哪儿有冰雪大世界景点评论 73032 条，其中大量的评论是用户未点评，系统默认的好评，采集这类数据没有意义，因此分别爬起来其中的好评、中评和差评，共 795 条。同程旅行爬取了从过去一直到 2024-02-22 的冰雪大世界景点所有在线评论，共 8089 条。

表 3.2 评论数据内容

日期	ip 属地	评分	评论
2024-02-22	江苏	1	非常差的体验，啥项目也没玩，花了一千多块钱，看了一堆快化的冰，不限流，也不退钱，简直抢钱！
2024-02-22	江苏	4	虽然去晚点气温升高了但冰雕好看
2024-02-22	上海	5	太值得一游，美轮美奂，北国风光。
2024-02-22	江苏	5	冰雪大世界非常棒，冰天雪地也是金山银山，点赞尔滨
2024-02-22	安徽	5	冰雕真的很震撼，雪地蹦野迪超带感
2024-02-21	云南	5	可以，很值得去的冰雪大世界，直接刷身份证就行。多穿点衣服，帽子耳罩围巾带点吃的也行。最好白天和晚上的风景都看一遍，各有不同
2024-02-21	福建	4	还行，就是去的时候冰雪开始融化了，一些东西玩不了

3.2.3 评论数据预处理

对爬取的评论数据进行简单的浏览，发现评论质量参差不齐，许多评论内容存在无法识别的字母、特殊符号等，又或者包含了大量重复或者无意义评论。通过对收集到的评论数据进行预处理，使文本数据更加干净、规范和易于处理，为后续的文本分析和挖掘提供更好的数据基础，保证了研究结果的准确性。

（1）数据清洗

对文本数据进行清洗处理，以去除不必要的信息、错误或噪声，从而提高数据质量和可用性。对原始数据进行以下几个步骤的清洗：

①使用 python 删除重复和空白数据，因为重复评论会导致数据重复，影响数据的准确性和可靠性，删除重复评论可以避免对分析和建模结果产生误导；

②用 Python 中的正则表达式 re 模块去除文本中的特殊字符，如 emoji 表情等特殊符号无法识别携带的语义信息，删除这些符号以减少噪音，使文本更加干净。部分评论去除符号后语义信息有所缺失，如‘冰雕艺术展👍’中的‘👍’删除后得到‘冰雕艺术展’就丢失了好评的情感语义，所以需要将其中的‘👍’转换成‘很棒’，然后将换号符转换成空格，多个逗号、句号、空格等合并为单个符号；

③用 python 的 opencc 库将评论中的繁体中文转换为简体中文，统一文本格式，减少数据处理的复杂性，提高文本处理的效率；

④游客的评论中出现多个连续相同的词语是一种常见的现象，这种情况可能是由于强调某种情感所导致的。例如，在“不错不错值得前去游玩”这句话中，重复的“不错”可能会导致情感分析算法将该评论判断为强烈的情感，但实际上该评论的情感并不一定那么强烈。针对这种情况，使用文本去重的方法来处理，帮助减少相同词语的重复出现，从而减少情感误导的可能性。

经过上述对在线评论进行数据清洗后，剩余有效评论 12162 条，清洗后的部分数据如下表 3.3 所示。

表 3.3 清洗前后评论数据对比示例

评论内容	清洗后评论
冰雪大世界适合下午游玩。最好多穿点衣服。晚上温度低。	冰雪大世界适合下午游玩。最好多穿点衣服。晚上温度低。
不错不错很方便	不错很方便
來哈爾濱必去景點，日間已經很美，晚上更美，就是排隊太花時間了	来哈尔滨必去景点，日间已经很美，晚上更美，就是排队太花时间了
哈尔滨得让人兴奋，借着女儿来哈尔滨比赛的机会，期待已久的旅行终于实现。	哈尔滨火得让人兴奋，借着女儿来哈尔滨比赛的机会，期待已久的旅行终于实现。

（2）分词

数据清洗之后需要进一步对评论进行分词，对文本数据进行分词是文本挖掘和分析中的重要步骤，能够帮助理解文本内容、提取关键信息和进行进一步的分析。目前常用的分词工具有 jieba 分词、THULAC、SnowNLP 等，其中最广泛应用分词工具为 jieba 分词，jjieba 分词支持用户自定义词典，根据需要添加专业名词、新词等，提高分词的准确性。本文使用 python 调用 jieba 包，在 jieba 词库基础上加入“冰雪大世界”、“雪圈”、“冰雪谷”、“暖宝宝”等冰雪景区专有名称词汇到词库当中，使用精确模式进行文本分词预处理。

表 3.4 评论分词示例

评论内容	分词
非常的好玩.非常漂亮.来哈尔滨必去的地方。非常推荐	非常 的 好玩 . 非常 漂亮 . 来 哈尔滨 必去 的 地方 。 非常 推荐
除了票价贵了，其他的都很好	除了 票价 贵 了 , 其他 的 都 很好
夜景很美，如身处童话世界中	夜景 很美 , 如 身处 童话世界 中
大滑梯和摩天轮的预约非常难，很难预约上。除非早上或者上午	大 滑梯 和 摩天轮 的 预约 非常 难 , 很难 预约 上 。 除非 早上 或者 上午

（3）停用词过滤

停用词是指在文本中频繁出现但通常对文本分析任务没有太大帮助的词语，

如“的”、“是”、“在”等一些常见的词语，它们在文本中频繁出现但往往没有太多实际含义，去除这些词语可以减少干扰，使得文本分析更加准确。本文使用哈工大停用词表，将词频较低且实际意义不大的连词、介词、数词及一些符号等删除，对冰雪大世界的评论分词进行过滤，处理后的结果如下表 3.5。

表 3.5 评论分词过滤示例

评论内容	分词
冰雪大世界很美丽哦	冰雪大世界 很 美丽
排队时间太长了，预约不太方便，现场二维码不好找，貌似故意不让人预约似的。	排队 时间 太长 预约 不太 方便 现场 二维码 不好 找 貌似 故意 不让 人 预约
不错的体验，就是人多	不错 体验 人多
真的很美，很壮观，但是大家一定要多穿衣服，小心冻感冒哦。	真的 很美 很 壮观 大家 一定 穿衣服 小心 冻 感冒

3.3 在线评论基本分析

3.3.1 描述性统计分析

（1）对采集的哈尔滨冰雪大世界评论数据进行了统计，共有 13386 条评论，其中 2015-2024 年的评论有 7482 条，评论数量随时间的变化如图 3.1 所示。

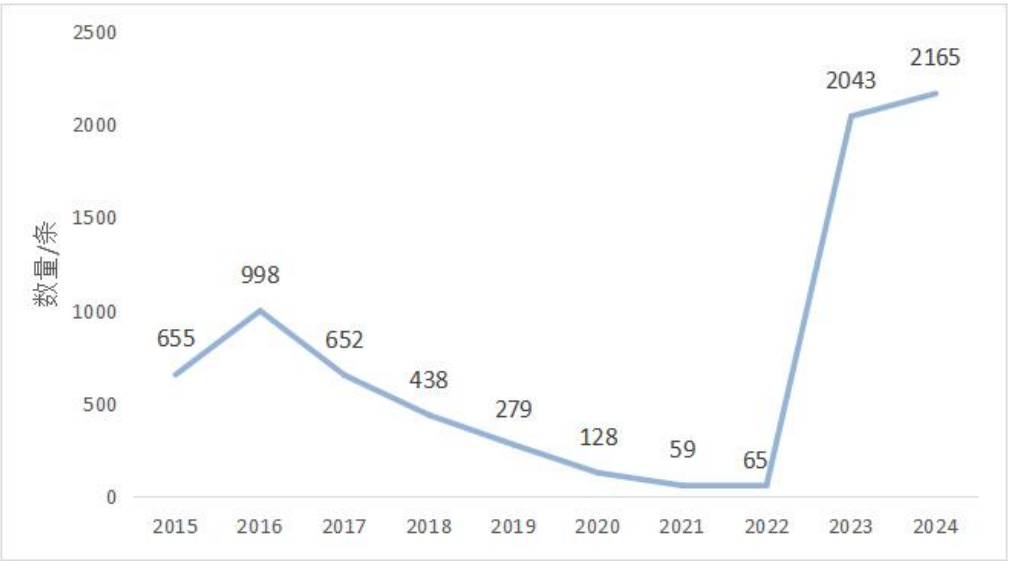

图 3.1 2015-2024 冰雪大世界评论数量

通过图 3.1 发现 2019-2022 年的评论数急剧下降，原因是新冠疫情对全球旅游业造成了巨大的冲击。然而，2022 年冰雪大世界景点的评论数量略微增加，说明随着疫情防控工作取得初步胜利，随着复工复产有序进行，人们的生活逐渐恢复正常，对于冰雪旅游的兴趣也在逐渐回升。直至 2022 年末疫情放开政策，旅游业逐渐恢复活力，人们对于冰雪旅游的热情重新被激发，因此 2023 年冰雪大世界的评论数急剧增长。2024 年冰雪大世界的评论数量比 2023 年要多，意味着随着疫情逐渐过去，人们对旅游的热情愈发高涨，尤其是对冰雪旅游的需求更加旺盛。

（2）采集的冰雪大世界评论中 2023 年和 2024 年的评论数量远高于其他年份，图 3.2 是 2023 年至 2024 年 2 月各月的评论数量。

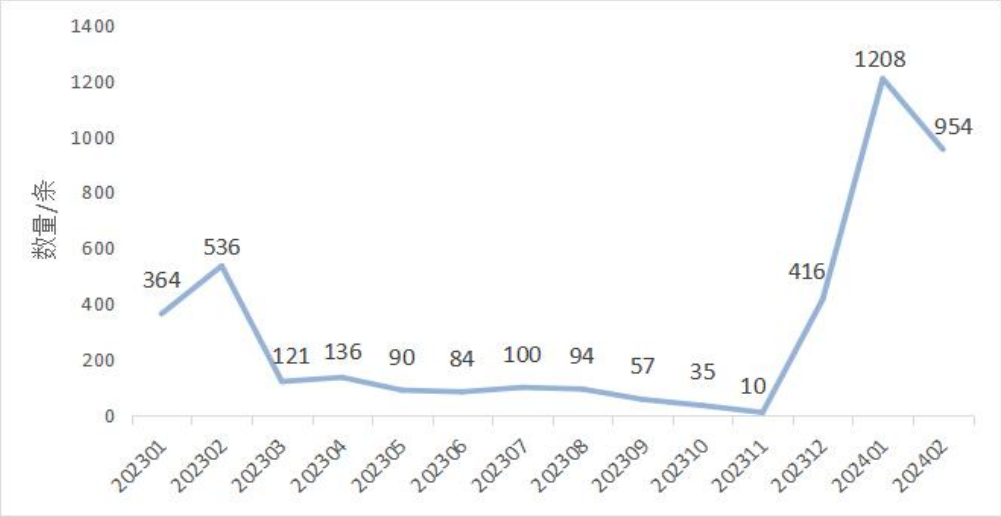

图 3.2　2023 年至 2024 年冰雪大世界评论数量

从图 3.2 中发现，冰雪大世界景点在一月和二月这两个月的游客评论数量最多，这是因与其特定的开放季节密切相关，哈尔滨冰雪大世界一般从 12 月中旬开启新一季的迎客序幕，一直到次年 2 月中旬结束，恰逢冬季寒冷时节，为游客提供了尽享冰雪娱乐的最佳时机。其中 2024 年一、二月份游客评论数量与 2023 年比显著增长，表明第 25 届冰雪大世界相较于上一届，推出了更多创新和吸引人的活动或项目，成功地吸引了更广泛的游客群体，从而让游客们更加积极地在线上分享他们的游览体验和感受。这一增长趋势不仅反映了冰雪大世界在活动创新方面的成功，也展示了冰雪旅游作为一种独特旅游形式的强大吸引力。

（3）第 25 届哈尔滨冰雪大世界开园时间是 2023 年 12 月、2024 年 1 月和 2024 年 2 月这三个月，共有评论 2578 条，其中好评、中评和差评的数量如下表 3.6 所示。

表 3.6 第 25 届冰雪大世界评论数量

评分	评论数量	占比(%)
好评	1973	76.53
中评	223	8.65
差评	382	14.82
合计	2578	100

由表 3.6 可知，采集的第 25 届哈尔滨冰雪大世界在线评论中好评和中评的数量加起来有 2196 条，占总评论的 85%以上。说明第 25 届哈尔滨冰雪大世界的举办取得了显著的成功，其丰富的游玩项目和独特的冰雪景观为大部分游客留下了深刻的印象和美好的回忆。但也有一部分游客的体验并不理想，对景区的评价较为负面，这些批评和建议是哈尔滨冰雪大世界在未来发展过程中需要重点关注和改进的地方。

3.3.2 词频分析

文本数据中的高频词通常反映出文本的重点内容，为了对哈尔滨冰雪大世界的评论数据进行直观了解，更好地认识游客对景点的点评，对文本数据进行词频统计。

（1）词云图分析

词云图是一种可视化工具，通过不同词语的大小和颜色来展示文本数据中词语的频率和重要性。在文本数据预处理的基础上，将冰雪大世界评论中所有词以词云图的形式可视化展示，把文本数据转化为视觉化的形式，使得数据更加直观、易于理解，结果如图 3.3 所示。

图 3.3 词云图

从图 3.3 中可以看出，词云图中的词语是游客在游玩冰雪大世界后的评论中出现频率较高的词语，即游客对冰雪大世界景点实际旅游体验的关注点和印象。“冰雪大世界”作为游客们游玩的旅游景点，自然会在他们的评论中频繁出现，成为词云图中最多的词语。“滑梯”、“冰雕”、“排队”、“景区”、“孩子”等这些关键词都是游客在冰雪大世界旅游体验中的重点关注和评价方向。

（2）词性分类分析

文本数据中不同词性的高频词所表示的含义也不尽相同，在描述哈尔滨冰雪大世界旅游品牌形象感知要素时，名词通常是最重要的要素，因为它能够直接表达出景点的核心特征和主题，形容词则描述旅游者在游览冰雪大世界时的感受和体验，动词可以描述旅游者在游览过程中所进行的行动和活动，进一步展现出旅游的乐趣和体验。因此，分别统计名词、动词、形容词的出现频率，得到词频 TOP10 的关键词如表 3.7 所示。

表 3.7 高频词表

名词	词频	形容词	词频	动词	词频
冰雪大世界	2934	方便	2303	排队	2572
滑梯	2521	非常	2103	取票	1623
项目	1824	不错	1697	表演	745
冰雕	1720	值得	1496	进去	656
哈尔滨	1613	没有	1440	游玩	639
里面	1465	很多	1148	服务	639
小时	1451	真的	1130	订票	638
晚上	1375	好玩	891	体验	533
门票	1145	开心	791	预约	516
同程	1135	特别	728	一去	455

高频词中占比最高的是名词，其中讨论最多的是“冰雪大世界”、“滑梯”、“项目”、“冰雕”、“哈尔滨”，这体现了冰雪大世界作为一个独特的冰雪主题公园，其景观和各种游玩项目是吸引游客的重要因素。形容词中词频较高的有“方便”、“非常”、“不错”、“值得”，反映了游客对冰雪大世界游玩的积极评价和和满意度。动词中词频最高的是“排队”、“取票”，说明冰雪大世界吸引了大量游客前来参观和体验，游客们需要排队等候才能购买门票或进入一些热门景点或项目。

3.4 LDA 主题分析

主题分析对挖掘游客评论具有重要作用，因为它可以快速了解评论的内容和重点，通过识别评论中的主题，更容易地理解游客们想要表达的情感信息和观点。在自然语言处理和文本挖掘领域，主题建模是一种常用的技术，用于从大量文本数据中自动发现和提取隐藏在其中的主题信息。其中，LDA 是一种常用的主题模型，它假设每个文档是由多个主题混合而成，每个主题又由一组词汇构成，通过对文档中的词汇分布进行建模，LDA 可以推断出文档中隐藏的主题结构。

对于冰雪大世界在线评论数据，经过清洗、分词等预处理后，为了更好地了解文本数据中的主题和情感，并进一步的挖掘游客对景区的关注点，对评论数据进行 LDA 主题模型的构建。将从多个旅游网站采集的在线评论划分类别，根据

游客在线评论打分，“5 分”和“4 分”等好评为正面情感类别，“2 分”和“1 分”等差评为负面情感类别，按照正面和负面两种情感类分别设置主题，其中正面情感类别有 9827 条评论，负面情感类别有 1271 条评论。本文选择文档-词频矩阵的方法分别对游客正负面情感评论文本建立 LDA 主题模型进行挖掘，提取并分析相关主题及主题的重点词，以此对景区管理、服务质量等提出相应对策，提升景区竞争力。

3.4.1 LDA 主题建模

使用基于词袋的 LDA 主题模型之前，需要准备好游客评论的语料库，创建词典和文档-词频矩阵。经过前文的预处理后得到了分词的评论数据，先筛选并删除数据中的单个字和频数小于 3 的词，然后剔除如“非常”、“真的”等无实际意义的词。因为单个字如“元”、“点”、“人”等作为主题中的主题词往往无法表达实际的语义信息，删除这些单个字可以减少对主题建模的影响；而频数小于 3 的词在数据集中出现的次数很少，这样的词很难提供足够的信息来帮助模型准确地学习主题结构。

分别创建正面情感评论和负面情感评论的词典和文档-词频矩阵，选择模型的参数，然后将创建的字典和文档-词频矩阵输入 LDA 主题模型中进行训练，模型的参数如表 3.8 所示。

表 3.8 LDA 主题模型参数

LDA 主题模型参数	数值
num_topics	5
passes	8
alpha	auto
eta	auto

其中，num_topics 是 LDA 主题模型的主题个数，passes 是训练过程中语料库的迭代次数、alpha 是文档-主题分布的先验，eta 是主题-词分布的先验，alpha=auto、eta=auto 意味着模型会自动学习并优化主题分布的稀疏性。

3.4.2 主题个数分析

评估 LDA 主题模型的常用方法是计算各个主题的困惑度和一致性。其中，困惑度是训练后获得的主题模型对某个文档所属主题的不确定性程度的不同，困惑度越低，表明不确定性越小，主题模型对主题聚类越有利。一致性是一种用于评估主题模型生成的主题质量的指标，评估主题之间的相关性和连贯性。本文采用困惑度和一致性作为确定最佳主题数目的指标，分别对正、负面情感评论选择了 1 到 20 不同的主题数量进行建模，困惑度和一致性随主题数目变化如图 3.4 和图 3.5 所示。

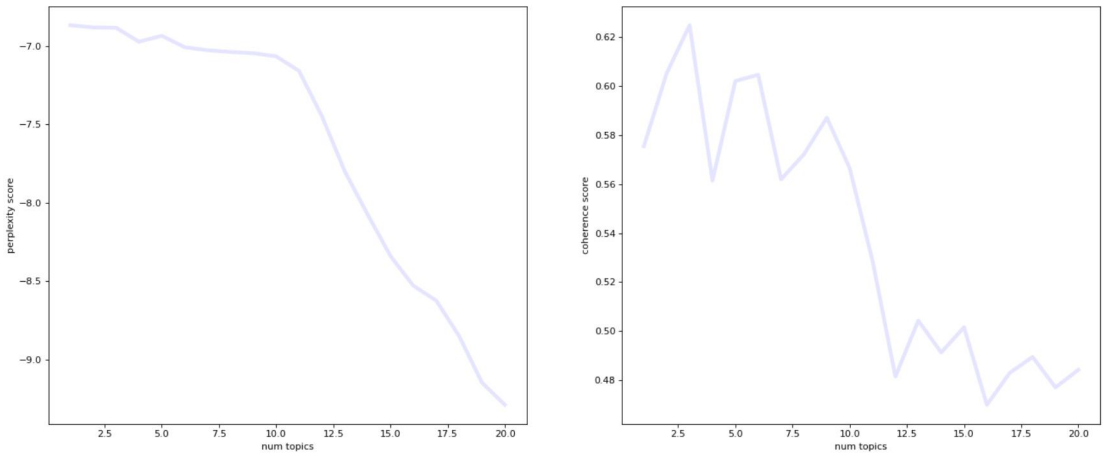

图 3.4 正面情感评论困惑度和一致性

由图 3.4 可知，正面情感评论的困惑度总体上随着主题数量的增加而缓慢下降，当主题个数超过 10 时，模型的困惑度下降速度加快，一致性随着主题数量的增加而波动变化。由困惑度的含义可知，困惑度越低越好，但并不是选择的主题数越多越好，因为当主题数过多时，模型会产生过拟合现象，所以需要根据一致性越大越好的原则确定主题个数，当主题个数达到 3 时，模型的一致性最大，因此，选择最终主题个数为 3。

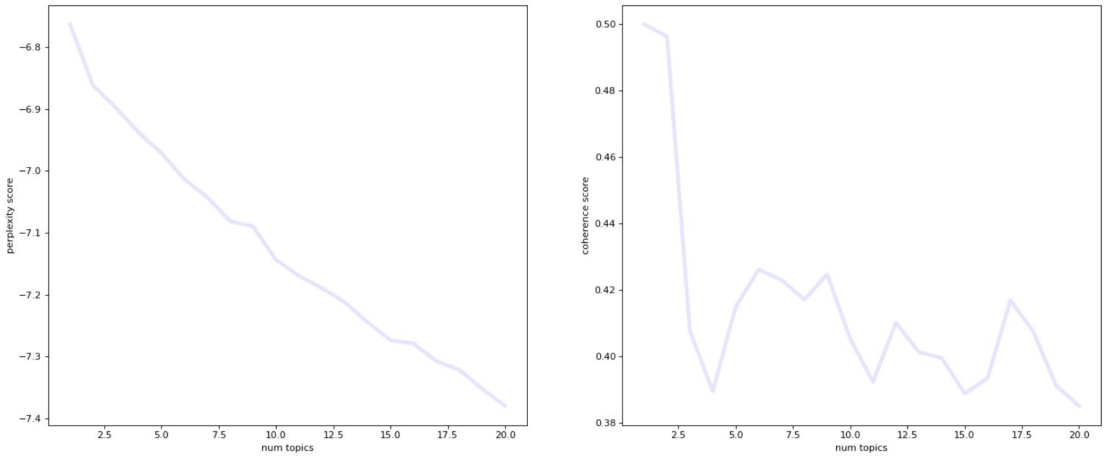

图 3.5 负正面情感评论困惑度和一致性

从图 3.5 发现，负面情感评论的困惑度随主题数目的增加而不断降低，一致性则随主题数目的增加波动变化，总体呈现下降的趋势，当主题个数为 2 时，模型的一致性较高，因此，最终确定负面评论的主题个数为 2。

四象限图是一种常用的可视化方法，将不同变量或主题分布在不同的象限中，以便更直观地比较和分析它们之间的关系。通过将文档-主题分布展示在坐标系中，可以清晰地看到不同主题在正、负评论中的分布情况，以及它们之间的关联和差异。主题挖掘的可视化结果如图 3.6 和图 3.7 所示。

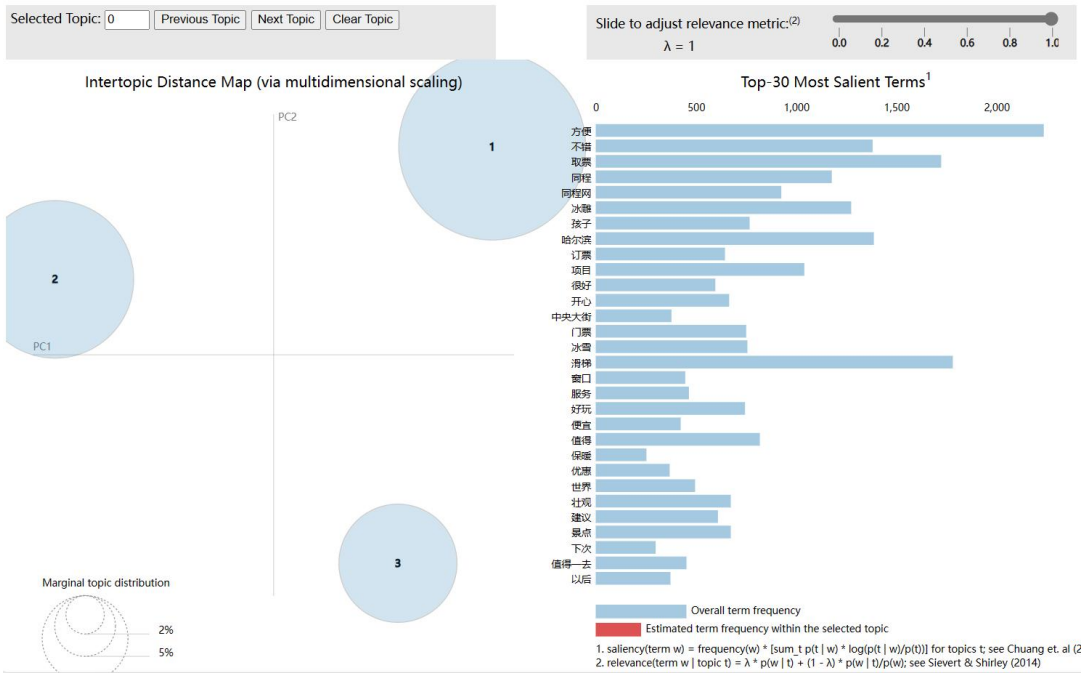

图 3.6 正面情感评论主题可视化图

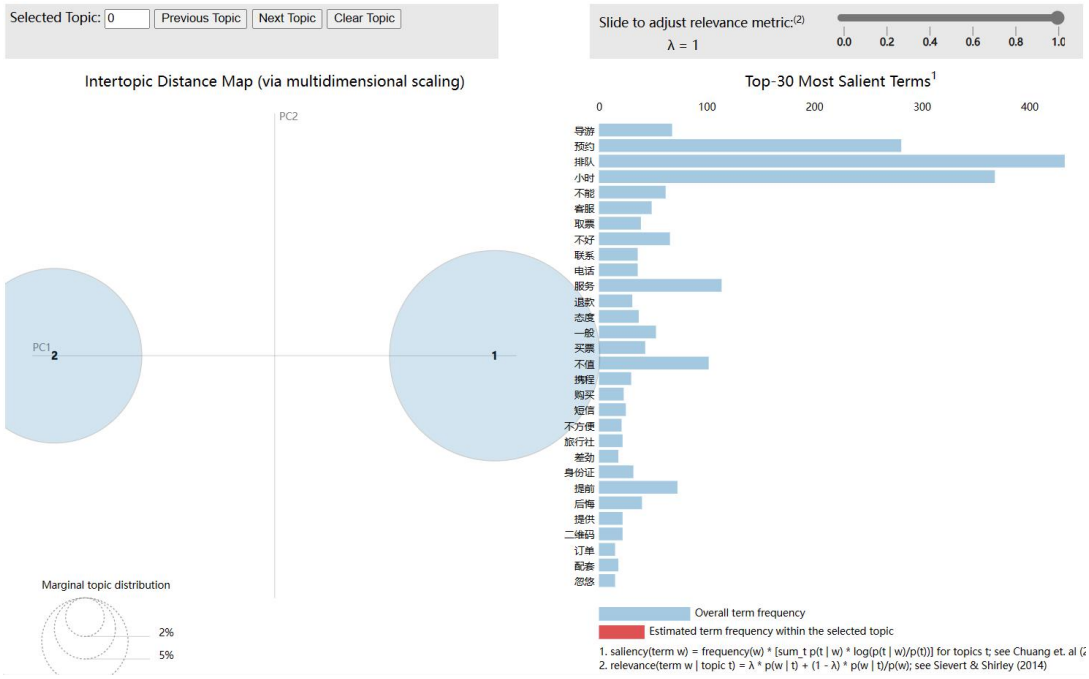

图 3.7 负面情感评论主题可视化图

在主题的可视化图中，通过观察气泡的大小和位置来判断主题的重要性和相似性，以及通过右侧的词频数量来了解每个主题的关键词。左侧的每个气泡代表一个主题，根据气泡的大小，直观地看出哪些主题在整体语料库中出现的频率较高，较大的气泡代表该主题在整体语料库中出现的频率较高，可能是比较普遍或具有主导地位的主题。气泡之间的距离代表主题之间的相似度，如果气泡之间存在重叠或者接近的现象，说明这两个或多个主题之间可能存在一定的关联或者交集，这种可视化方式可以快速识别出主题之间的相似性和关联性。

图 3.6 和图 3.7 中的气泡互不重叠，且分布比较均匀，表示所建立的主题模型是合理的，这种情况下，比较容易区分和理解每个主题所代表的内容，而且主题之间的关联性较低，不会造成混淆，提高主题模型的可解释性和有效性。

3.4.3 主题结果分析

经过主题提取后，分别输出正、负面情感评论每个主题及其对应的主题词，各个主题后对应的 TOP20 关键词如表 3.9 和表 3.10 所示：

表 3.9 正面情感评论主题-主题词

Topic1	Topic2	Topic3	Topic1	Topic2	Topic3
方便	冰雪大世界	中央大街	排队	好玩	马迭尔
取票	滑梯	冰雪大世界	景点	壮观	打车
不错	哈尔滨	保暖	景区	漂亮	宾馆
同程	冰雕	建议	服务	很漂亮	公交车
同程网	项目	时间	窗口	表演	消费
孩子	晚上	预定	便宜	小时	注意
门票	排队	出来	优惠	世界	小时
订票	冰灯	门口	以后	很多	出租车
开心	值得	下午	冰雪大世界	很美	表演
很好	冰雪	公交	有点	值得一去	排队

表 3.9 是正面情感评论的主题特征词，部分主题词存在一定程度的重叠，这是因为不同主题之间可能会共享一些相似的词语或主题特征，这种情况下，即使主题词有一定的重叠，也不会影响对评论主题的讨论和理解。对三个主题分析总结如下：

主题一是服务与价格体验，从“方便”、“同程网”、“订票”等这些词语表明游客关注的是冰雪大世界购票的便利性，在线购票服务让游客更加方便快捷地购买门票，避免了排队等待的麻烦。此外，“取票”、“服务”、“窗口”等词说明购票服务的质量和窗口服务也是游客非常关心的问题，因为良好的服务可以提升游客的体验和满意度。最后，提到了“门票”、“便宜”、“优惠”等词，说明门票价格对游客来说也是一个重要的考量因素，门票价格的优惠活动也会吸引更多游客前来体验。

主题二是景点体验，从“滑梯”、“冰雕”、“值得”、“好玩”、“值得一去”等词看出冰雪大世界的游玩项目确实给游客们留下了深刻的印象，滑梯和冰雕等项目似乎给他们带来了很多乐趣和惊喜，让他们感到非常满意和满足。如“壮观”、“漂亮”、“很漂亮”、“很美”等词，游客对冰雪大世界的景观赞美有加，认为这些景观非常壮观、美丽，这些赞美之词表明这些冰雪景点的美感和震撼力给他们留下了深刻的印象。

主题三是旅游体验，从“建议”、“注意”、“保暖”等词可知，游客在讨论冰雪大世界的游玩体验时，可能会提出一些建议和注意的问题，以帮助其他人更好地享受游玩过程。“门口”、“公交”、“打车”、“宾馆”、“公交车”、“出租车”等词说明游客关注的是景区的交通问题，对于那些选择乘坐公共交通工具的游客来说，他们会关注如何方便地乘坐公交车、出租车等交通工具，以便更好地游览景区。“中央大街”、“马迭尔”等词说明游客对于冰雪大世界当地的特色景点和品牌比较关注，中央大街和马迭尔等地方为游客提供了独特的风景和体验，让他们感受到了当地的文化和魅力。

表 3.10 负面情感评论主题-主题词

Topic1	Topic2	Topic1	Topic2
门票	排队	门口	时间
服务	小时	取票	收费
导游	预约	买票	差评
排队	门票	联系	体验感
不好	工作人员	电话	不值得
小时	体验	园区	票价
不能	不值	以后	景点
客服	很多	垃圾	入园
体验	提前	态度	分钟
一般	游玩	不让	建议

表 3.10 是负面情感评论的主题特征词，对两个主题分析总结如下：

主题一是服务体验，从“服务”、“导游”、“客服”、“联系”、“电话”、“态度”等词可知，游客对于导游、客服的景区服务并不满意，如果导游的服务不到位或者态度不好，会给游客留下不好的印象，影响到他们对景区的整体评价。“不好”、“不能”、“一般”、“垃圾”、“不让”等词是游客们对于冰雪大世界的评价，看出他们对冰雪大世界的体验并不理想。

主题二是游玩体验，从“排队”、“小时”、“预约”、“提前”、“时间”、“分钟”等词看出游客们对于冰雪大世界的排队等候体验并不满意，长时间排队会影响游

客的游玩体验，导致不愉快的情绪。“门票”、“不值”、“收费”、“票价”看出游客们对于冰雪大世界的门票价格和景区的收费标准有所不满，门票价格过高或者与景区提供的服务和体验不相符可能会让游客感到不值得。

通过对哈尔滨冰雪大世界的在线评论进行了 LDA 主题分析，深入地了解游客对景点的不同方面的评价和意见，看出游客对哈尔滨冰雪大世界景区的购票便利性、景区的游玩观赏性、交通便利性以及当地特色美食方面都非常满意，这些方面可能是吸引游客前来的重要因素，让他们在景区内度过愉快的时光。然而，部分游客对景区的服务质量和人多排队等方面表示不满，这意味着景区在管理和服务方面还有一些改进的空间，可以考虑增加服务人员数量、优化排队系统等措施来提升游客体验。

4 冰雪大世界在线评论情感分析

数字旅游时代的到来，游客习惯于在各类平台上分享他们的旅行经历和观点，通过挖掘和分析这些评论文本，对其进行情感分析帮助了解游客对于景点的情感倾向和态度。目前的情感分析方法有基于情感词典、传统机器学习和深度学习的方法，情感词典的方法通过对情感词汇进行统计和打分判断文本的情感倾向，传统机器学习方法主要利用特征工程和分类器来进行情感分析，而深度学习方法则通过构建深度神经网络模型来自动学习文本的特征表示和情感分类。

本章对从旅游网站采集到的冰雪大世界在线评论数据进行情感分析，分别从情感词典、传统机器学习和深度学习出发进行分析。然后提出一种 BERT-EW 双通道情感分析模型，先使用情感词典提取评论文本的情感词特征，然后基于 BERT 构建双通道模型将文本的情感词特征与语义特征进行融合，从而突出评论的情感信息，提高情感分析的准确性和效果。

4.1 基于情感词典的情感分析

早期，大多数情感分析基于语义规则的情感词典方式，它使用预定义的情感词典（也称为情感词库或情感词汇表）来评估文本的情绪倾向。在情感词典中，词汇的情感强度或权重通常是通过人工评估或基于大规模语料库的统计方法得出的，如果一个词汇在正面文本中出现的频率超过在负面文本中的频率，那么这个词可能会被赋予一个正面的情感强度。

本文将基于情感词典进行整合和扩充，通过结合一定的评分规则对评论进行情感打分。首先，构造一个情感词典，收录与情感相关的词汇，如正面情感词汇“喜欢”、“满意”等、负面情感词汇“讨厌”、“失望”等。然后，根据评论中出现的情感词汇，结合一定的评分规则，对评论进行情感打分。

4.1.1 情感词典的构建

（1）基础情感词典

目前最常用的词典有清华大学李军中文褒贬义词典、Hownet 知网情感词典、台湾大学 NTUSD 和大连理工大学中文情感词汇本体库。情感词典通常分为四个部分，积极情感词典、消极情感词典、否定词典以及程度副词词典。积极情感词典和消极情感词典分别收录了表达正面和负面情绪的词汇，如“喜欢”代表积极情

绪，“讨厌”则归于消极情绪。否定词词典包含能改变句子情感倾向的否定词，如“不”、“没”等，这些词汇在分析时会对情感的判断产生重要影响。程度副词词典则涵盖了能够强化或弱化情感表达强度的副词，如“非常”、“稍微”等，这类词汇对于精准捕捉情感强度至关重要。

为了提高文本情感分析的准确度，整合和扩充基础情感词典是一个关键步骤，这一过程涉及将多个来源的情感词汇合并，同时去除重复项，形成一个更全面、更精确的情感词典。通过将 Hownet 知网的正面评价词语、NTUSD 的积极词典和清华大学李军中文褒义词典等的正面词汇合并，去除重复后，得到一个基础的积极情感词典；同理，通过合并得到一个基础的消极情感词典。然后将分词后的情感词、否定词和程度副词通过情感词典文件、否定词文件、程度副词文件分别放入三个字典，为下面计算情感分数做铺垫。

（2）扩充情感词典

通用的情感词典虽然在文本情感分析中起到了重要作用，但由于不同领域的评论对象和语境存在差异，一些在特定领域具有明显情感倾向的词汇可能在通用情感词典中并未包含。因此，针对特定领域的个性化需求，需要对情感词典进行扩充和优化，在这个过程中，采用基于 Word2Vec 的词义相似度计算方法，来识别和添加语义上接近的词汇。

首先需要确定一组具有明显情感倾向的种子词，这些种子词需要根据具体的领域和语境进行自定义选择。理想情况下，从整个语料库中找出具有强烈情感倾向的词汇作为种子词，但这在海量的文本中可能会耗费大量的工作量。因此，本文借助 TF-IDF（Term Frequency-Inverse Document Frequency）算法从文本中提取关键词作为情感种子词，TF-IDF 是一种统计方法，它能够反映出词汇在文本中的重要程度，从而从大量的文本中快速有效地筛选出具有代表性的情感种子词。

对预处理后的评论文本进行了 TF-IDF 计算，根据计算出的 TF-IDF 值对词进行了排序，并选择了排名前 1000 的关键词。这些关键词是基于它们在文本中的重要性选择的，因此，它们很可能包含了评论中的主要观点和情感倾向，TOP20 的关键词如表 4.1 所示：

表 4.1 TF-IDF 前 20 关键词

关键词			
好好	冰雕	回忆	童话世界
一如既往	使用方便	不错	尔滨
太冷	期待	震撼	人太多
开心	很好	冰雪奇缘	永远
值得	滑梯	身份证	排队

借助 TF-IDF 所输出的关键词中，人工选取其中具有明显情感倾向的词语作为正负面情感种子词，其中选择正面情感词语和负面情感词语各 100 个，部分情感种子词如表 4.2 所示：

表 4.2 情感种子词

正面		负面	
好好	冰雪奇缘	太冷	无语
一如既往	童话世界	人太多	千万别
开心	好玩	排队	不值
值得	值得一看	好久	不好
使用方便	干净	排不上	退钱
期待	环境优美	毛病	没建好
很好	没白来	辛苦	不值得
回忆	巧夺天工	不好玩	太贵
不错	快捷	差差	坐不上
震撼	强烈推荐	垃圾	费劲

使用 Word2Vec 模型计算其他词语与情感种子词的相似度，对于每个非种子词语，找出与其最相似的种子词，并计算相似度分数。设置相似度阈值为 0.6，只有当非种子词与种子词的相似度超过阈值时，才考虑将其添加到情感词典中，为新添加的词赋予与其最相似的种子词相同的情感倾向。最后得到一个扩充后的情感词典，它不仅包含了原来的情感种子词，还包含了一些与情感种子词具有高

度相似性的词语，从而提高了情感词典的覆盖率和分析的准确性。

4.1.2 评分规则

设定一个简单的情感词语权重系统，假设情感值遵循线性叠加原理，每个正面情感词语的情感值被赋予+1，每个负面情感词语被赋予-1，在进行句子分词后，如果词语向量包含在情感词典中，就将其对应的值加入到句子的总情感分数中。需要注意的是，否定词和程度副词具有特殊的处理规则，对于否定词，其出现会导致紧接着的情感词语权重反号，而对于程度副词，则根据其强化或者减弱的程度，将其乘以一个相应的系数，在本文中，设定了六个档次的修饰程度，分别对应[0.5，0.75，1.2，1.5，2，3]的权重。此外，还需要检查句子的结束标点，如果一个正面情感词语的分句结尾有感叹号，那么其情感值将增加 2，而对于负面情感词语，如果其分句结尾有问号或感叹号，那么其情感值将减少 2。最后，将所有的情感分数相加，得到最终的情感得分，根据这个得分的正负，判断评论文本的整体情感倾向，评分规则如图 4.1 所示。

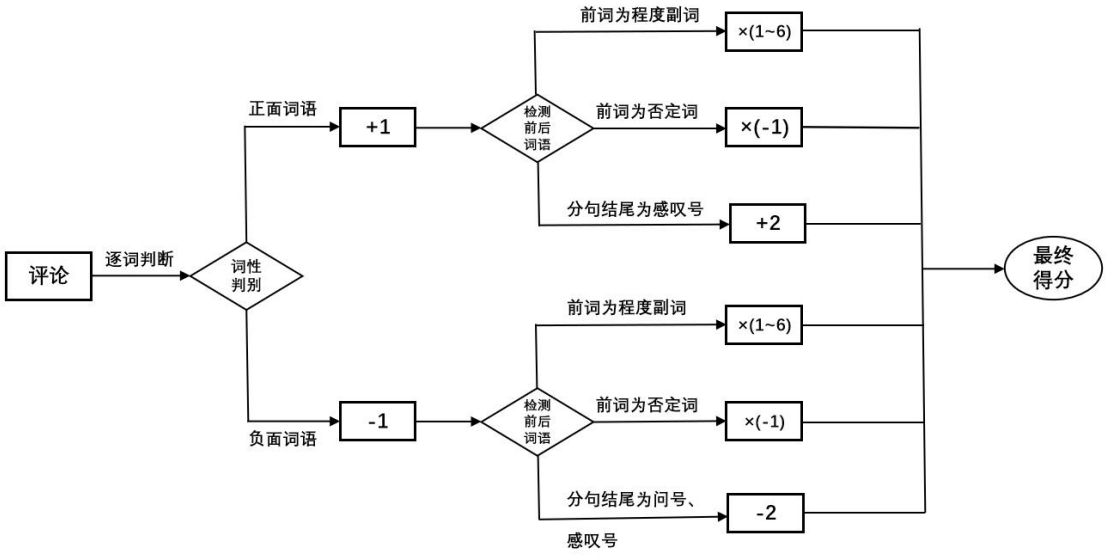

图 4.1 情感词典评分规则

4.1.3 实验数据与评价指标

（1）实验数据

本次实验使用的数据是从多个旅游网站采集的哈尔滨冰雪大世界游客在线评论，清洗后得到共 12162 条差评、中评和好评三种情感极性的评论数据，选择

其中的好评和差评作为情感分析的实验数据，数据具体情况如表 4.3 所示。

表 4.3 实验数据统计表

情感类别	数量	占比(%)
正面	9827	88.55
负面	1271	11.45
总和	11098	100

从表 4.3 中发现，数据集中好评的数量几乎是差评数量的 8 倍，说明数据集存在着类别不平衡情况，而数据的不均衡问题往往会让模型更偏向于多数类的样本，对少数类样本的识别表现不佳。

（2）评价指标

情感分析任务是自然语言处理中常见的任务之一，评估模型在情感分析中的表现通常会使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F1 值等指标。

准确率（Accuracy）是所有预测中正确预测的比例。准确率计算公式为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4-1}$$

精准率（Precision）是指在模型预测为正例的情况下，有多少样本真正是正例，精准率衡量了模型的准确性，精准率越高，说明模型在预测正例时的准确性越高。精准率的计算公式为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4-2}$$

召回率（Recall）又称为查全率，表示所有真正正例中，有多少被模型正确识别为正例，衡量了模型对正例的识别能力。召回率的计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4-3}$$

F1 值是一个综合考虑了模型的查准率和查全率的指标，是精确率和召回率的调和平均值，能够更全面地评估模型的性能。F1 值的计算公式为：

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4-4}$$

其中，TP（True Positive）表示正样本被模型预测为正样本的个数，TN（True Negative）表示负样本被预测为负样本的个数，FP（False Positive）表示负样本被模型预测为正样本的个数，FN（False Negative）表示正样本被模型预测为负样本的个数。

值得注意的是，召回率和精确率是两个相互影响的指标，它们之间存在一种权衡关系，通常希望模型能够在保持较高的精确率的同时，也能有较高的召回率，即尽可能减少误判的同时尽可能减少漏判。因此，单独使用精确率或召回率无法全面准确地评价模型的分类能力，而 F1 值是精确率和召回率的调和均值，能够平衡精确率和召回率之间的关系，通过它可以更全面地评价模型的分类效果。

4.1.4 结果分析

基于情感词典的情感分析结果如表 4.4 所示：

表 4.4 情感分析结果

情感类别	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
负面	44.48	59.64	50.96	86.85
正面	94.54	90.37	92.41	

结果表明，正面评论的 F1 达到 92.41%，在识别正面评论方面表现良好，大部分正面评论都被准确地识别出来，而且误报率较低，负面评论方面的 F1 只有 50.96%，识别效果较差。这些结果可能受到选择的情感词典，以及设定的评分规则等因素的影响，为了提高情感分析的效果，需要进一步扩充和优化情感词典，以及调整评分规则。此外，对于一些复杂的情感表达和隐含的情感信息，可能还需要引入更复杂的分析方法，如深度学习等。

4.2 基于传统机器学习的情感分析

基于传统机器学习的情感分析方法，主要依赖于从文本中提取有效的特征，然后应用各种统计和机器学习算法对这些特征进行训练和分类。这些传统的机器学习方法在处理结构化和半结构化数据时表现出色，能够在大量的文本数据中找出有用的模式和关联。然而，这些方法也有其局限性，比如难以处理非结构化文本、依赖于人工特征工程、对大规模数据的处理效率低下等。尽管如此，基于传

统机器学习的情感分析仍然是情感分析研究的重要组成部分，它提供了理解和分析文本情感的有力工具，也为后续的深度学习和神经网络方法提供了基础和启示。

4.2.1 特征提取

对于预处理后的评论，将文本转换为机器学习模型可识别的数值特征，词袋模型（Bag of Words，BoW）和词频-逆文档频率（TF-IDF）是两种常见的文本特征提取方法。特别地，TF-IDF 是一种在文本挖掘和信息检索领域广泛应用的技术，旨在评估一个词语在文档集合或语料库中某份文档的重要性。其核心思想是如果一个词语在某篇文档中频繁出现，但在其他文档中较少见，那么这个词语对于区分该文档内容具有较高的价值。

采用 TF-IDF 方法提取文本特征，并将这些文本转换成特征向量，使其能够被传统的机器学习算法如朴素贝叶斯、支持向量机、决策树等进行有效的情感分析处理。TF-IDF 的计算公式如(4-5)-(4-7)：

$$词频(TF) = \frac{词W在文档中出现的次数}{文档的总词数} \tag{4-5}$$

$$逆文档频率(IDF) = \log(\frac{文档总数}{包含词W的文档数}) \tag{4-6}$$

$$TF - IDF = TF * IDF \tag{4-7}$$

4.2.2 模型训练

在本研究中，采用了 Sklearn 机器学习库来实现情感分析，通过决策树、逻辑回归、支持向量机以及朴素贝叶斯四种不同的机器学习算法对比分析了各模型的分类效果。实验数据被分为训练集和测试集，其中训练集用于模型训练，测试集用于评估模型的预测性能，将实验数据按照 4:1 的比例划分为训练集和测试集。

机器学习模型的性能不仅受限于算法的选择，还极大依赖于超参数的设置，超参数的优化对于提高模型的分类性能至关重要，理想的超参数设置能够显著提升模型性能。网格搜索（Grid Search）通过遍历指定的参数范围内所有可能的参数组合，为每一种组合构建模型，并评估其性能，从而选出最优的模型设置。这种方法虽然计算量较大，但能够系统地寻找到最佳参数组合，确保获得的分类模型达到最优性能。

为了获得最佳的实验结果，本文采用了网格搜索方法对机器学习模型的超参数进行了优化，以下是各个机器学习模型中所设定的参数。

（1）决策树参数

表 4.5 决策树参数网格

模型	参数	数值
决策树	criterion	['gini', 'entropy']
	max_depth	[None, 10, 20, 30]
	min_samples_split	[2, 10, 50]
	min_samples_leaf	[1, 5, 10]
	max_features	[None, 'auto', 'sqrt', 'log2']
	max_leaf_nodes	[None, 5, 10, 20]

其中，criterion 是用来测量分裂质量的函数，常用的有 gini（基尼不纯度）和 entropy（信息增益），max_depth 为树的最大深度，min_samples_split 表示一个节点必须具有至少这么多的样本才能分裂，min_samples_leaf 表示叶子节点必须有至少这么多的样本，这个参数限制树的生长，max_features 为寻找最佳分割时考虑的最大特征数量，可以是整数、浮点数、字符串或 None，max_leaf_nodes 为最大叶子节点数。

（2）逻辑回归参数

表 4.6 逻辑回归参数网格

模型	参数	数值
逻辑回归	C	[0.01, 0.1, 1, 10, 100]
	penalty	['l1', 'l2']
	solver	['liblinear', 'saga']

其中，C 表示正则化强度的逆，较小的值指定更强的正则化，penalty 参数用于指定用于正则化的范式，常见选项有 l1 和 l2，solver 参数指定了在优化问题中使用的算法。

（3）支持向量机参数

表 4.7 支持向量机参数网格

模型	参数	数值
支持向量机	C	[0.001, 0.01, 0.1, 1, 10, 100]
	kernel	['linear', 'rbf', 'poly']
	gamma	[0.0001, 0.001, 0.01, 0.1, 10]

其中，C 是正则化参数，其值为正数，较小的 C 使得决策边界更平滑，而较大的 C 旨在正确分类所有训练样本，可能会导致模型过拟合，kernel 参数指定了在算法中使用的核函数类型，可以是 linear、poly、rbf、sigmoid 等，gamma 定义了单个训练样本影响的范围，较大的值意味着更接近的样本有更强的影响。

（4）多项式朴素贝叶斯参数

表 4.8 多项式朴素贝叶斯参数网格

模型	参数	数值
多项式朴素贝叶斯	alpha	[0.01, 0.1, 0.5, 1, 2, 10]
	fit_prior	[True, False]

其中，alpha 是拉普拉斯或利德斯通平滑参数，此参数用于防止零概率问题，默认值为 1.0，即拉普拉斯平滑，fit_prior 是一个布尔参数，表示是否学习类别的先验概率，如果为 False，则所有类别的先验概率都设为等于。

4.2.3 结果分析

使用不同算法在训练集上训练分类模型，借助网格搜索交叉验证的方式寻找最优超参数，然后把测试集的数据代入到模型中加以验证，四种算法所训练出的分类模型评价指标如表 4.9 所示：

表 4.9 情感分析结果

模型	评论类别	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
决策树	负面	60.09	55.24	57.56	90.90
	正面	94.43	95.39	94.90	
逻辑回归	负面	77.51	65.32	70.90	94.01
	正面	95.72	97.62	96.66	
支持向量机	负面	80.00	64.52	71.43	94.23
	正面	95.64	97.97	96.79	
朴素贝叶斯	负面	81.59	66.13	73.05	94.55
	正面	95.84	98.12	96.97	

结果展示，朴素贝叶斯表现最佳，正面评论和负面评论的 F1 分别为 96.97%和 73.05%，展示出了较高的准确率和 F1 得分。逻辑回归和 SVM 的表现相近，决策树在处理正面评论时效果很好，但在负面评论的处理上显著不如其他算法，尤其是在精确率和召回率上，这是由于决策树模型较易过拟合的特性所致。

4.3 基于深度学习的情感分析

传统的情感分析方法主要基于词典或者机器学习，词典方法依赖于预定义的情感词典，而机器学习方法则需要大量人工标注的数据。然而，这些方法在处理口语化、错别字、网络用语等非正式文本时效果较差，对于语境的理解和长距离依赖关系的捕捉能力也较弱。

深度学习以其独特的优势，为情感分析带来了新的可能，深度学习能够自动学习并抽取文本特征，对于处理非正式文本和理解语境有着较强的能力。尤其是一些特定的深度学习模型，如循环神经网络（RNN）、长短期记忆网络（LSTM）和变压器（Transformer），它们对于捕捉文本的长距离依赖关系具有较强的能力，这对于情感分析任务来说尤为重要。因此，基于深度学习的情感分析已经成为当前研究的热点和趋势，许多研究工作和应用系统都在尝试利用深度学习技术来提升情感分析的性能。

4.3.1 词嵌入模型

词嵌入模型是自然语言处理中的一种核心技术，它能够将文本中的单词或短

语转换为固定长度的密集向量。这些向量代表着词汇的语义信息，使得机器学习模型能够更好地理解和处理自然语言数据。词嵌入模型的目标是捕获单词之间的语义和语法关系，将这些关系编码到一个多维空间中，其中语义或语法相似的词汇被映射到相近的点。

目前，最为流行的词嵌入模型之一是 Word2Vec，它由 Google 研发并于 2013 年公开发布，Word2Vec 基于神经网络模型，能够通过学习大规模语料库中的词汇共现信息来生成词向量。Word2Vec 模型主要包括两种训练架构，连续词袋（continuous bag of words，CBOW）和 Skip-Gram，CBOW 预测目标单词基于上下文，而 Skip-Gram 则正好相反，它预测上下文基于目标单词。

除了 Word2Vec，GloVe 和 FastText 也是两种广受欢迎的词嵌入模型，各自带来了新的视角和优势。GloVe 模型侧重于单词共现的全局统计信息，它结合了矩阵分解技术和窗口上下文方法的优点，通过对整个语料库的共现矩阵进行操作，生成词向量。FastText 模型则引入了子词信息，特别是在处理形态丰富的语言时显示出了优势。

4.3.2 BERT 预训练模型

BERT 是 Google 的 Devlin 等于 2018 年 10 月提出的预训练语言模型，一举刷新了 11 个 NLP 任务的榜单。该模型基于 Transformer 的双向编码器表示，是一个预训练的语言表征模型，它强调了不再像以往一样采用传统的单向语言模型或者把两个单向语言模型进行浅层拼接的方法进行预训练，而是采用新的 MLM（Masked Language Model）和 NSP（Next Sentence Prediction）两种任务进行预训练，以致能生成深度的双向语言表征。

（1）模型结构

BERT 模型的结构是基于 Transformer 的 Encoder，主要由多个 Transformer 模块堆叠而成。结构如图 4.2 所示，其中 E1，E2，…，EN 表示字的文本输入，输入文本会经过字嵌入、段嵌入和位置嵌入三部分的处理，然后通过多层 Transformer 编码器进行处理，最终得到文本的向量化表示。

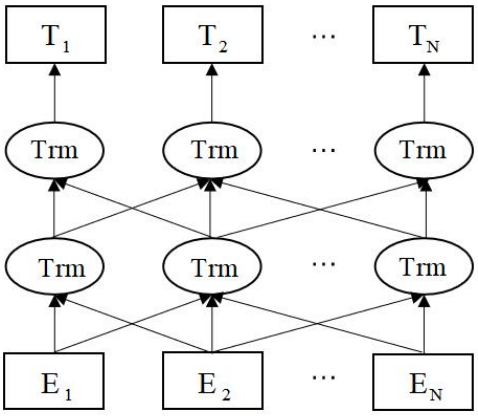

图 4.2 BERT 模型结构

（2）模型输入

图 4.3 是将文本中的每个字转换为向量作为 BERT 模型输入，由图可知模型输入嵌入由字向量、位置向量、分段向量三部分相加得到。BERT 的输入部分是个文本序列，序列的开头、结尾会添加一个特殊的[CLS]、[SEP]，表示整个序列的开始和结束。每个单词有三个 embedding，位置信息 embedding，对单词顺序进行编码，单词 embedding，将文本中的每个字转换为词向量，句子 embedding，区分训练数据的两个句子，把每个字对应的三个 embedding 叠加，就形成了 BERT 的输入。

输入	[CLS]	明	年	还	去	[SEP]	加	油	尔	滨	[SEP]
词向量向量	E[CLS]	E明	E年	E还	E去	E[SEP]	E加	E油	E尔	E滨	E[SEP]
段向量	EA	EA	EA	EA	EA	EA	EB	EB	EB	EB	EB
位置向量	E0	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10

图 4.3 BERT 模型输入

（3）模型输出

BERT 模型的输出有两种形式，一种是字符级别的向量，对于输入序列中的每个词，BERT 模型会生成一个对应的词向量表示，经过多层 Transformer 编码器的处理，每个词的词向量会包含该词在上下文中丰富的语义信息；另一种是句子级别的向量，BERT 模型会对输入序列中的特殊标记“[CLS]”进行处理，对应

“[CLS]”标记的词向量表示被用作整个句子的表示，这个句子级别的向量经过训练，能够捕捉整个句子的语义信息和上下文关系，如图 4.4 所示。

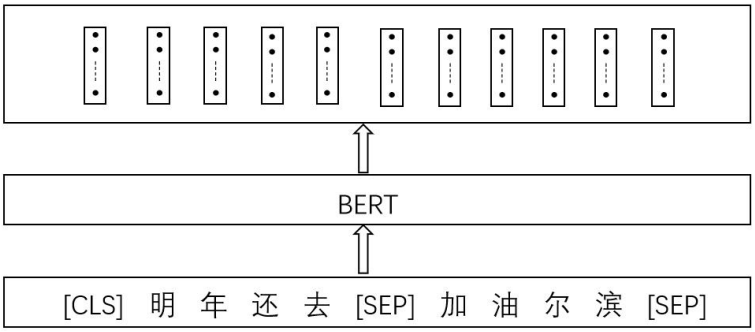

图 4.4 BERT 模型输出

4.3.3 模型训练

使用深度学习框架 Pytorch 训练模型，评估和比较不同深度学习模型在情感分析任务上的性能。在模型设计上，采用 Word2Vec 模型将文本数据转换成词向量，随后将这些词向量作为输入，分别接入到 CNN、BiGRU 以及 BiLSTM 三种不同的神经网络模型中。此外，还引入了基于 Transformer 架构的 BERT 预训练模型，使用中文的 BERT-base 模型。

为确保所有模型在相同的训练条件下进行，将一系列相关参数保持一致。设置文本截取长度以覆盖 80%的文本，其中，Word2Vec 模型的最大截取长度为 30，而 BERT 模型为 80，训练批次设置为 32，迭代次数为 10 次，在优化方面，Word2Vec 模型的学习率设定为 1e-4，而 BERT 模型的学习率为 2e-5，使用 Adam 算法作为网络优化器，损失函数则采用交叉熵损失函数。

为了保证各个模型能在情感分析任务上取得良好的性能，对于 CNN 神经网络，进行了两次卷积池化操作，对于 BiGRU 和 BiLSTM，使用网格搜索的方法寻找最优的模型参数，具体的参数设置如下所示。

（1）CNN 参数

表 4.10 CNN 参数

模型	参数	数值
CNN	filter_sizes	3
	num_filters1	200
	output_size1	25
	num_filters2	100
	output_size2	20

其中，filter_sizes 为卷积核的尺寸，代表卷积核覆盖的单词数，num_filters 表示每种尺寸的卷积核的数量，决定了卷积层的输出通道数，output_size 表示自适应平均池化输出的长度。

（2）BiGRU 和 BiLSTM 参数

表 4.11 循环神经网络参数

模型	参数	数值
BiGRU、BiLSTM	hidden_size	[64, 128, 256]
	num_layers	[1, 2, 3, 4, 5]

其中，hidden_size 为隐藏层的特征数量，决定了隐藏状态的维度，num_layers 为神经网络堆叠的层数。

4.3.4 结果分析

对比了使用 Word2Vec 词向量结合不同深度学习模型和直接使用 BERT 模型在情感分类任务上的表现，结果显示，这些模型在负面和正面评论分类任务上均展示了不错的性能，具体表现如表 4.12 所示。

表 4.12 情感分析结果

模型	评论类别	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
W2V-TextCNN	负面	76.64	75.40	76.02	94.69
	正面	96.91	97.11	97.01	
W2V-BiGRU	负面	80.09	74.60	77.24	95.09
	正面	96.83	97.67	97.25	
W2V-BiLSTM	负面	82.27	72.98	77.35	95.23
	正面	96.65	98.02	97.33	
BERT	负面	83.87	83.87	83.87	96.40
	正面	97.97	97.97	97.97	

从表中可以看出，当使用正面情感评论作为评价标准，各模型的 F1 值均达到 97%以上，显示出了较高的性能和较小的性能差异，因此使用负面情感评论来评价模型的分类效果。比较前三个模型，发现使用 Word2Vec 词向量的 TextCNN、BiGRU 和 BiLSTM 的 F1 值分别为 76.02%、77.24%和 77.35%，BiLSTM 和 BiGRU 模型的 F1 值要高于 TextCNN，说明双向循环神经网络在提取文本特征上的效果要优于卷积神经网络。对比采用动态词向量生成方式的 BERT 模型与基于静态词向量 Word2Vec 的神经网络模型，BERT 模型在所有评估指标上都表现出了优越性，这凸显了预训练模型在深层次理解语言结构和含义方面的强大能力，这得益于其深层的双向结构，使其能够更全面地理解文本上下文，进而在情感分类任务上达到更高的准确率。

4.4 基于 BERT-EW 的双通道情感分析模型

4.4.1 BERT-EW 双通道模型

本文在 BERT 模型基础上融合评论文本的情感词特征与语义特征搭建了 BERT-EW 双通道情感分析模型。如图 4.5 所示，它主要由三个部分组成，分别是语义特征提取、情感词特征提取和输出分类。

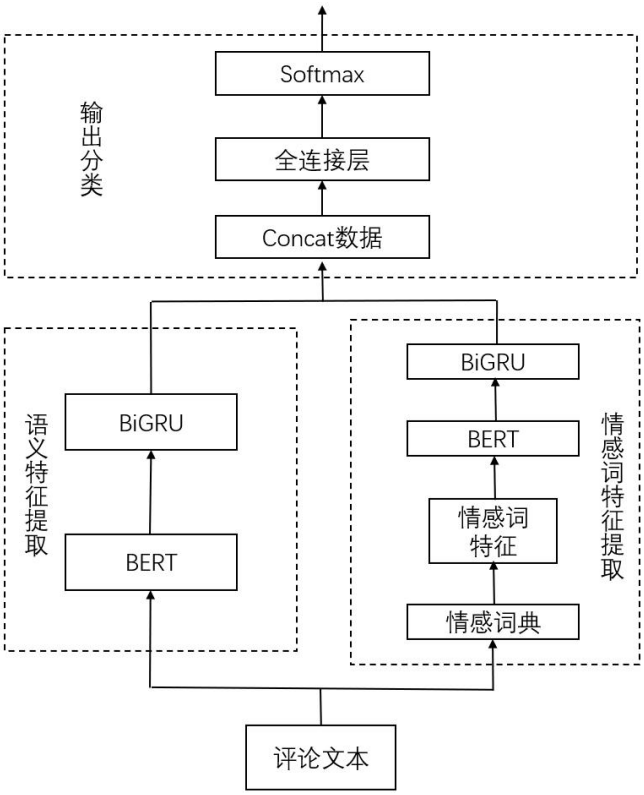

图 4.5 BERT-EW 双通道模型结构图

（1）语义特征提取

评论文本先转换成 BERT 模型可接收和处理的特定格式向量，由文本的词索引、遮掩标识和句子的分段标识组成，本实验是在 BERT 模型的基础上进行情感分析，不需要考虑两个句子之间的关系，所以输入向量只需要词索引、遮掩标识即可。

将文本向量输入 BERT 模型得到每个字在上下文中的动态词向量，BERT 由 12 层 Transformer 组成，融合 BERT 模型的多层特征，将 BERT 模型最后四层的输出相加得到最终的文本语义向量表示W，计算公式如(4-8)-(4-9)所示：

$$T_i = (x_{i1}, x_{i2}, x_{i3}, \cdots, x_{in}) \tag{4-8}$$

$$W = \sum_{i=9}^{12} T_i \tag{4-9}$$

其中，n 表示文本的长度，x_{ij} 表示第 j 个字的向量，T_i 是 BERT 模型的第 i 层输出。

将蕴含语意的特征向量W输入 BiGRU 网络，进一步挖掘评论数据的上下文

语义信息，得到文中更丰富和深层的语义特征向量G，计算公式如(4-10)-(4-13)所示：

$$\overrightarrow{g}_t = GRU(x_t, \overrightarrow{g}_{t-1}) \tag{4-10}$$

$$\overleftarrow{g}_t = GRU(x_t, \overleftarrow{g}_{t-1}) \tag{4-11}$$

$$g_t = w_1 \overrightarrow{g}_t + w_2 \overleftarrow{g}_t \tag{4-12}$$

$$G = (g_1, g_2, g_3, \cdots, g_n) \tag{4-13}$$

其中，$\overrightarrow{g}_t$为由前向后在 t 时刻的隐藏层状态，$\overleftarrow{g}_t$为由后向前在 t 时刻的隐藏层状态，$\overrightarrow{g}_t$和$\overleftarrow{g}_t$加权求和得到 BiGRU 在 t 时刻的隐藏层状态g_t，G为 BiGRU 输出的特征向量。

（2）情感词特征提取

通过情感词典提取评论的情感词，将每条评论的情感词组成情感特征文本，例如评论“也是南方小土豆，第二次带孩子去了，孩子很喜欢，冰雪王国，很梦幻”，提取出情感词“南方小土豆”、“很”、“喜欢”、“王国”、“梦幻”，得到情感词特征文本“南方小土豆很喜欢王国很梦幻”。将由情感词构成的文本输入 BERT 模型得到多层输出融合的情感词特征向量$E = (x_1, x_2, x_3, \cdots, x_m)$，然后使用 BiGRU 网络对情感词向量$E$进行特征提取，得到整个文本的情感词特征表示$G'$。

（3）输出分类

将文本的语义特征向量和情感词特征向量进行拼接之后得到向量C，然后将向量C输入到全连接层中，通过 Softmax 分类函数输出最终的分类概率 S，继而得到到文本的情感类别，计算公式如(4-14)-(4-17)所示：

$$C = G \oplus G' \tag{4-14}$$

$$H = \text{Relu}(W_h C + b_h) \tag{4-15}$$

$$D = W_d H + b_d \tag{4-16}$$

$$S = \text{Softmax}(D) \tag{4-17}$$

其中，G语义特征向量，G'为情感词特征向量，H、D为全连接层输出，W

和b分别为权重和偏置。

4.4.2 参数设置

设置 BERT-EW 模型的相关参数，评论文本长度为 80，情感词特征长度为 30，训练批次设置为 32，迭代次数为 10 次，学习率设定为 2e-5，采用了 Adam 优化器和交叉熵损失函数。对于 BiGRU 神经网络，采用了网格搜索的方法来进行参数调优，具体的参数值如表 4.13 所示。

表 4.13 BiGRU 神经网络参数

参数	数值
hidden_size	[64, 128, 256]
num_layers1	[1, 2, 3, 4, 5]
num_layers2	[1, 2, 3, 4, 5]

其中，hidden_size 为隐藏层的特征数量，num_layers1 为提取语义特征的神经网络层数，num_layers2 为提取情感词特征的神经网络层数。

4.4.3 结果分析

得到了 BERT-EW 双通道情感分析模型的效果，并进一步与 BERT 和 BERT-BiGRU 模型的性能进行了比较，结果展示在表 4.14 中。

表 4.14 情感分析结果

模型	评论类别	Precision(%)	Recall(%)	F1(%)	Accuracy(%)
BERT	负面	83.87	83.87	83.87	96.40
	正面	97.97	97.97	97.97	
BERT-BiGRU	负面	84	84.68	84.34	96.49
	正面	98.07	97.97	98.02	
BERT-EW	负面	81.62	89.52	85.38	96.58
	正面	98.67	97.46	98.06	

BERT-EW 模型采用了双通道处理机制，分别处理评论的语义特征和情感词特征，在这种双通道架构下，语义通道和情感词通道的输出将被合并，形成一个

综合特征表示，这有助于模型捕获更丰富和更精确的情感信息。从表中可以看出，BERT 模型后接入 BiGRU 比单一的 BERT 模型效果更好，说明在 BERT 的输出中加入 BiGRU 能够提取句子中更深层次的情感特征，从而提升分类准确性。BERT-EW 的正面评论和负面评论的 F1 分别为 98.06%和 85.38%，与 BERT-BiGRU 相比，BERT-EW 在 F1 和 Accuracy 指标上的表现要优于 BERT-BiGRU 模型，展示了优秀的性能。这种将语义特征和情感词特征融合的策略，突出了文本中的情感信息，使得模型能够更准确地理解和判断文本的情感倾向，BERT-EW 模型在处理包含复杂和细微情感变化的文本数据时，具有很大的优势和潜力。

5 总结与展望

5.1 结论

随着我国旅游业的飞速发展，其中的冰雪旅游在今年更是成为备受瞩目的主题，其热度居高不下。冰雪旅游在各地都备受关注，吸引了大量游客前来体验冰雪乐趣，为当地经济和旅游业发展带来了新的机遇和挑战，因此了解游客的需求和情感反馈尤为重要。本文以哈尔滨冰雪大世界冰雪旅游景点为例，采集旅游网站上的冰雪大世界在线评论数据，对数据进行主题挖掘和情感分析，得出如下研究结论：

（1）通过 LDA 主题模型了解到，冰雪大世界的游客正面情感主要倾向景点体验、服务与价格体验和旅游体验三个主题，游客在景点体验方面主要关注的是冰雪大世界景点的观赏性，服务与价格体验上的关注点是订票、取票的便利和门票的价格及优惠，旅游体验中游客比较关注景区的交通、当地特色及保暖等内容。游客负面情感主要倾向服务体验和游玩体验两个主题，服务体验中游客不满意的主要是景区导游、客服的服务，在游玩体验方面对冰雪大世界感到不满意的是排队时间太长、门票价格和景区的收费太贵等。

（2）在进行情感分析时，研究发现传统机器学习方法相较于依赖情感词典的方法表现更佳，尤其是采用多项式朴素贝叶斯算法时，分析效果要优于其他传统机器学习模型。进一步地，使用深度学习技术对情感分析进行增强时，结果表明效果有显著提升，特别是在运用 BERT 模型进行哈尔滨冰雪大世界评论的情感分析时，与传统词嵌入模型相比，模型的准确率实现了大幅度的提升。这说明深度学习模型，尤其是 BERT 模型，在处理复杂的语言情感分析任务时，能够提供更高的准确性和效率。

（3）本研究提出的 BER-EW 双通道情感模型结合了评论文本的语义信息与情感词信息，显著提高了对哈尔滨冰雪大世界评论的分类准确性，其效果相较于 BERT-BiGRU 模型更为显著。该模型的成功应用不仅证明了其在情感分析领域的优越性能，也为管理者和业内人士提供了一种高效的工具，以深入理解游客的情感反馈，进而优化服务质量和游客体验。

5.2 建议

对于冰雪大世界的在线评论进行了文本挖掘，使用 LDA 主题模型进行主题分析，根据分析的结果提出以下几种提升冰雪大世界市场竞争力的建议：

（1）主题分析显示游客对冰雪大世界的冰雪景观和娱乐项目的需求较高，可以考虑建造更多样化的冰雕景观，设计不同主题的冰雕展区，如童话世界、历史文化、自然景观等，打造可以互动的冰雕景观，如冰雕迷宫、冰雕滑梯；增加丰富的娱乐项目，如滑雪、雪地摩托车、雪地探险和冰雪游乐园等。

（2）针对景区门票价格和收费高昂的问题，景区可以将提供的基础服务与增值服务进行分离，为游客提供多样化的选择；根据游客需求和服务内容，设定不同档次的门票价格，如成人票、学生票、儿童票等，满足不同人群的需求；争取政府对景区的税收减免政策，减轻景区的财政压力，降低门票价格。

（3）针对景区导游、客服的服务问题，对冰雪大世界的售票员、导游、酒店从业人员等进行多方面的培训，提高他们的工作态度和服务水平。可以通过以下措施，制定系统的培训计划、建立持续的培训和考核机制、优化工作环境和福利待遇、建立客户反馈和改进机制。

（4）针对游客排队时间太长的问题，冰雪大世界可以提高管理水平，实时监控景区内的游客流量和安全情况，建立高效的清洁和维护机制；还可以开发智能排队系统，游客通过智能排队系统随时查看排队人数，并根据游客的需求和偏好，智能规划最佳游览路线。

5.3 不足与展望

（1）本文的数据选取来自于携程、同程、去哪儿三个旅游平台上的游客评论，采集的评论数据不够全面，对于冰雪大世界的在线评论分析有一定的影响。为了更全面地了解冰雪大世界的在线评论情况，可以考虑从其他平台或渠道获取更多的评论数据，比如在微博等社交媒体上搜索相关话题或关键词，收集游客的反馈和评论，获得更多来源的评论数据，从不同角度和渠道了解游客对冰雪大世界的评价，有助于更全面地分析和理解游客的情感倾向。

（2）本文对冰雪大世界在线评论的情感是根据游客的打分划分为好评、中评和差评，可能存在少数游客的评价内容和评分有所差异，所以在进行情感分析时，可能会对模型的准确性产生影响。在情感分析之前，可以增加预处理步骤来

识别并处理那些评分与评论内容明显不匹配的数据，然后进行适当的调整或标注。

参考文献

[1] Alsumait L,Barbara D,Domeniconi C.On-line LDA:Adaptive topic models for mining text streams with applications topic detection and tracing//Proceedings of the 8th IEEE International Conference on Data Mining.Pisa,Italy,2008:3-12.

[2] Blei D M,Ng A Y,Jordan M I.Latent Dirichlet Allocation[J].Journal of Machine Learning Research,2003,3(4/5): 993-1022.

[3] Cui Y M,Che W X,Liu T,et al.Revisiting pre-trained models for chinese natural language processing[C].Conference on Empirical Methods in Natural Language Processing (EMNLP)2020:657-668.

[4] Cui Y,Che W,Wang S,et al.LERT:a linguistically-motivated pre-trained language model[EB/OL].arXiv:2211.05344,2022.

[5] Dong Z,Dong Q.HowNet-a hybrid language and knowledge resource[C]//International Conference on Natural Language Processing and Knowledge engineering,2003. Proceedings.2003.IEEE,2003:820-824.

[6] Das R,Zaheer M,Dyer C.Gaussian LDA for topic models with word embeddings// Processing of the Annual Meeting of the Association for Computational Linguistics and the Join Conference on Natural Language Processing of the Asian Federation of Natural Language Processing. Beijing, China,2015:795-804.

[7] Dey R,Salem F M.Gate-variants of gated recurrent unit(GRU)neural networks [C]//2017 IEEE 60th Inter national Midwest Symposium on Circuits and System(MWSCAS),2017 :1597-1600.

[8] Devlin J,Chang M W,Lee K,et al.BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J].arXiv preprint arXiv:1810.04805,2018.

[9] Fang B,Ye Q,Kucukusta D,et al. Analysis of the perceived value of onli ne tourism reviews:Influence of readability and reviewer characteristics[J].Tourism Management, 2016,52(2):498-506.

[10] Hochreietr S,Schmidhuber J.Long short-term memory[J].Neural Computation,1997,9:17 35-1780.

[11] Ku L,Lo Y,Chen H.Using polarity scores of words for sentence-level opinion Extraction

[C]//Proceedings of NTCIR-6 workshop meeting.2007:316-322.

[12] Liu Y H,Ott M,Goyal N,et al.BoBERTa:A Robustly Optimized BERT Pretraining Approach[OL].arXiv Preprint,arXiv:1907.11692.

[13] Lecun Y,Bottou L,Bengio Y,et al.Gradient-based learning applied to document recognition[J].Proceedings of the IEEE,1998,86(11):2278-2324.

[14] Lan Z,Chen M,Goodman S,et al. Albert:A lite bert for self-supervised learning of language representations[J].arXiv preprint arXiv:1909.11942,2019.

[15] Mikolov T,Sutskever I,Chen K,et al.Distributed Representations of Words and Phrases an Their Compositionality[C].Advances in Neural Information Processing Systems,2013: 3111-3119.

[16] Mikolov T,Chen K,Corrado G,et al.Efficient estimation of word representations in vector space[J].arXiv preprint arXiv:1301.3781,2013.

[17] Pang B,Lee L,Vaithyanathan S.Thumbs up?Sentiment classification using machine learning techniques[J].arXiv:cs/0205070,2002.

[18] Tang D,Qin B,Liu T.Document modeling with gated current neural network for sentient classification[C]//Proceedings of the 2015 Conference on empirical methods in natural language processing.2015:1422-1432.

[19] Torresen E N,Singh D,Robertsonrinf A.Consumer reviews and the creation of booking transaction value:Lessons from the hotel industry[J].International Journal of Hospitality Management,2015(50):77-83.

[20] Williams D,Hintom G.Learning representations by back-propagating errors.Nature, 1986,323(6088):533-538.

[21] Wu Meifen,Long Ruyin,Chen Feiyu,Chen Hong,Bai Yun,Cheng Kun,Huang Han. Spatio-temporal difference analysis in climate change topics and sentiment orientation: Based on LDA and BiLSTM model[J].Resources, Conservation & Recycling,2023,188.

[22] Xu L,Lin H,Pan Y,et al.Constructing the affective lexicon ontology[J].Journal of the China society for scientific and technical information,2008,27(2):180-185.

[23] Zhang Z,Han X,Liu Z,et al.ERNIE:enhanced language representation with informative entities[C]//Proceedings of the 57th annual meeting of the association for computational linguistics.Florence:ACL,2019:1441-1451.

[24] 段丹丹,唐加山,温勇,袁克海.基于 BERT 模型的中文短文本分类算法[J].计算机工程,2021,47 (01):79-86.

[25] 吉兴全,曾若梅,张玉敏,宋峰,孙鹏凯,赵国航.基于注意力机制的 CNN-LSTM 短期电价预测[J].电力系统保护与控制,2022,50(17):25-132.

[26] 胡荣磊,芮璐,齐筱,等.基于循环神经网络和注意力模型的文本情感分析[J].计算机应用研究,2019,36(11):3282-3285.

[27] 姜兆国.哈市冰雪大世界景点评论的文本挖掘分析[D].广西师范大学,2023.

[28] 刘龙飞,等.基于卷积神经网络的微博情感倾向性分析[J].中文信息学报,2015,29 (6):159-165.

[29] 李彦冬,郝宗波,雷航.卷积神经网络研究综述[J].计算机应用,2016,36(09):2508-2515+2565.

[30] 刘逸,保继刚,陈凯琪.中国赴澳大利亚游客的情感特征研究——基于大数据的文本分析[J].旅游学刊,2017,32(05):46-58.

[31] 刘思琴,冯胥睿瑞.基于 BERT 的文本情感分析[J].信息安全研究,2020,6(03):220-227.

[32] 李磊,吴旭辉,刘继.融合关键对象识别与深层自注意力的 BiLSTM 情感分析模型[J].小型微型计算机系统,2021,42(3):504-509.

[33] 刘干,林杰豪,翟雯熠.基于中心词和 LDA 的微博热点话题发现研究[J].情报杂志,2021,40(05):143-148+164.

[34] 刘经纬,张淑琪.基于情感分析的微博热点话题演化分析[J].信息系统工程,2022,(12):137-140.

[35] 陆之洲.平台经济背景下在线评论对销量的影响研究[D].上海财经大学,2023.

[36] 綦方中,田宇阳.基于 BERT 和 LDA 模型的酒店评论文本挖掘[J].计算机应用与软件,2023,40(07):71-76+90.

[37] 司育.基于 LDA 主题模型的山西省热门景区综合评价体系研究[D].山西财经大学,2024.

[38] 唐慧丰,谭松波,程学旗.基于监督学习的中文情感分类技术比较研究[J].中文信息学报,2007,21(6):88-94.

[39] 魏泽阳.基于 BERT 的中文文本情感分析研究[D].西安电子科技大学,2022.

[40] 王承云,戴添乐,蒋世敏等.基于网络大数据的上海红色旅游形象感知与情感评价研究[J].旅游科学,2022,36(02):138-150.

[41] 王晨,廖启明.基于改进的 LDA 模型的文献主题挖掘与演化趋势研究——以个人隐私信息保护领域为例[J].情报科学,2023,41(10):112-120.

[42] 肖红,许少华.基于句法分析和情感词典的网络舆情倾向性分析研究[J].小型微型计算机系统,2014,35(04):811-813.

[43] 涂海丽,唐晓波.基于在线评论的游客情感分析模型构建[J].现代情报,2016,36(04):70-77.

[44] 荀竹.基于 LERT 和双通道模型的微博评论情感分析研究[J].计算机时代,2023,(10):80-82+88.

[45] 严仲培,陆文星,束柬等.面向旅游在线评论情感词典构建方法[J].计算机应用研究,2019,36(06):1660-1664.

[46] 游兰,曾晗,韩凡宇等.基于 BERT-BiGRU 集成学习的情感语义识别[J].计算机技术与发展,2023,33(05):159-166.

[47] 阮光册,黄韵莹.融合 Sentence-BERT 和 LDA 的评论文本主题识别[J].现代情报,2023,43(05):46-53.

[48] 张晨逸,孙建伶,丁轶群.基于 MB-LDA 模型的微博主题挖掘[J].计算机研究与发展,2011,48(10):1795-1802.

[49] 张仰森,等.基于双重注意力模型的微博情感分析方法[J].清华大学学报:自然科学版,2018,58(2):122-130.

[50] 赵宏,傅兆阳,赵凡.基于 BERT 和层次化 Attention 的微博情感分析研究[J].计算机工程与应用,2022,58(05):156-162.

[51] 赵金雨.黄山风景区在线评论的游客情感分析[D].哈尔滨商业大学,2022.

[52] 诸林云,曲金帅,范菁等.基于 BERT-BiLSTM-Attention 的文本情感分析[J].云南民族大学学报(自然科学版),2023:1-11.

[53] 张一彤.基于主题建模与情感分析的网络舆情研究[D].山西财经大学,2023.

致 谢

光阴似箭，日月如梭，三年的硕士研究生求学即将结束，回想期间的学习和生活，面对培育我的母校，心中无限感慨。在完成本篇硕士学位论文之际，我要向所有在我研究过程中给予帮助和支持的人们表示最诚挚的感谢。

首先，我要衷心感谢我的导师刘明教授，在整个研究过程中，他给予了我无私的指导和悉心的关怀。他的深厚的学术造诣和严谨的治学态度对我产生了深远的影响，他的指导和鼓励使我能够克服困难，不断进步，我也要感谢他在论文写作和研究方法上的指导，使我能够更好地完成研究工作。

我亦感激我的同学和朋友们，感谢他们在过去的几年里的陪伴和支持，我们共同度过了许多难忘的时光，他们的鼓励和理解是我完成学业的重要动力。

此外，我要感谢我的家人和朋友们，他们在我整个研究生阶段给予了我无私的支持和鼓励，他们的理解和支持使我能够专注于研究工作，克服了各种困难。

最后，我要感谢我的母校兰州财经大学，在这里，我接受了优质的教育和培养，为我今后的发展奠定了坚实的基础。我也要感谢学校提供的各种资源和设施，为我的研究工作提供了便利条件。

在此，我要再次向所有给予我帮助和支持的人们致以最诚挚的祝福，没有你们的支持，我将无法完成这篇论文。谢谢！

徐应发

2024 年 3 月