北京外国语大学

毕　业　论　文

题　　目：跨语言新闻事件情感演化问题研究

学　　院：**计算机科学与技术学院**

专　　业：**计算机科学与技术**

姓　　名：**徐墨馨**

攻 读 学 位：**工学学士**

导　　师：**梁野**

定 稿 日 期：**2020** 年 **05** 月 **14** 日

Research on Emotional Evolution of Cross-Language News Events

Dissertation Submitted to

Beijing Foreign Studies University

in partial fulfillment of the requirement

for the degree of

B.E. in Computer Science

By

Moxin Xu

(Computer Science & Technology)

Dissertation Supervisor : **Ye Liang**

May, 2020

声　明

本人郑重声明：

所呈交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的研究成果。若论文查重重复率超过 20%，本人自愿重写毕业论文；

若所提交毕业论文的电子版与纸版论文内容不一致，本人自愿放弃答辩资格；

所提交毕业论文终稿已经过导师审阅并符合毕业论文基本要求，导师同意并推荐参加论文答辩。

论文作者（签名）：　　导师（签名）：

日期：　　日期：

摘要

随着互联网技术的迅猛发展，新闻网站和社交媒体逐渐取代纸质报刊成为人们获取新闻的主要来源。由于网络传播的即时性，不同国家的新闻事件也能实时更新同步。基于上述背景，跨语言新闻事件情感演化问题的研究实现批量查询含有关键词的不同国家新闻和在时间维度上对新闻文本进行情感倾向分析，从而挖掘跨语言新闻传播时间节点，掌握舆情趋势。

本篇论文首先在第一章绪论中详细介绍了跨语言新闻演化问题研究背景，现状及意义，对该研究涉及的技术进行了简要介绍。对于运用的两大关键技术，网络爬虫技术和情感分析技术，第二章和第三章详细介绍了它们的研究成果和设计架构。第四章阐述了利用前两章技术实际研究泰国和马来西亚两国新闻演化问题的过程。第五章总结实践中遇到的一些问题和收获并对未来研究趋势作出展望。

本次研究参考了国内外数十篇有关网络爬虫、情感分析、舆情系统、机器学习等领域论文，对论文中涉及的有关算法进行了阐述、分析和实践。实践操作时获取了泰国和马来西亚两个国家的各两个新闻网站新闻，将单篇新闻链接、标题、发布时间和内容保存到本地 excel 文件中，并建立情感分析模型对于两国新闻文本情感进行分析，最后将数据批量上传到本地数据库。本地系统环境为 Windows10，编程语言使用 Python3,数据库为本地 MYSQL 数据库，访问数据库软件为 Navicat。实际操作过程中各个模块均能够正常运行，实现了研究的预期目标。

关键词：跨语言新闻事件，情感分析，网络爬虫

Abstract

With the rapid development of Internet technology, news websites and social media have gradually replaced newspapers, magazines and periodicals as people's main source of news. Due to the immediacy of network transmission, the news events that happened in different countries can also be uploaded and synchronized on the internet at the same time. Therefore, different countries' net users can share the information and get to know what's happening around the world. Based on the background, the research on the emotional evolution of cross-language news events is proposed and applied to analyze the emotion propensity of people in different countries to the same topic. This paper contains two important steps of the study, firstly crawling batch of different national news from international websites and secondly analyzing the sentiment of crawled news texts, so as to establish a database with news and sentiment. Finally, the users could find all the available international news with the same key word, understand the process of message transmission and analyze the sentiment of different countries toward the same issue.

This paper first introduces the background, current status, and significance of the study of the evolution of cross-language news event in detail, and briefly introduces the technologies involved in the study in Chapter One. For the applied two key technologies, namely web crawler technology and sentiment analysis technology, Chapters Two and Three explain their algorithm and up-to-date research results collected from academic journals and papers. The fourth chapter takes the techniques introduced in the above two chapters into practice, in order to study the sentimental evolution of news in Thailand and Malaysia. The last chapter summarizes some problems and accomplishments during the above study process, as well as makes plans for future research.

The research refers to a number of domestic and foreign papers concerning many fields such as web crawlers, sentiment analysis, public opinion systems, machine learning, artificial intelligence, natural language processing and so on. What's more, it covers the explanation, analysis, comparison and realization of some related complicated algorithms and discusses the feasibility in the reality, which are comprehensive and enlightening. In the operating part, the Thailand and Malaysia news from two websites respectively are saved to local directory with crawling machine. With the format of link, title, release time and content of a single news article, the data are crawled and saved in excel documents with the name of website number and collecting date. Sentiment analysis models for Thailand and Malaysia are established accordingly to analyze the news from the two countries. Finally the data are uploaded to the local database in batches. Users could search for news by typing in the key word in the database. The result would be shown in the order of the publish time of the news with their web page links, titles and sentiments, positive or negative.

All the study are based on the following configuration. The local system environment is Windows 10, the programming language is Python3, the database is a local MYSQL database, and the software accessed to database is Navicat. During the actual operation, each module can run normally, and the expected goal of the research is achieved.

Key words: Cross-language news events, sentiment analysis, web crawlers

目录

第 1 章　绪论

1.1　研究背景及意义

目前随着网络的发展以及科技的进步，新闻平台比如新浪网凤凰网已经超越纸质报刊成为人们获取新闻的途径。这些平台提供的国际新闻大多数实时更新，具有即时性的特点，正好作为新闻事件采集的来源。人们目前看到的国际新闻都是跨语言传播的，即新闻发源国发布报道后，其他国家用其他语言转述报道，转述过程中报道的角度和情感态度或多或少会发生变化，这些情感倾向会直接影响受众的态度，进而影响国际关系。在瞬息万变的国际局势中，中国作为正在崛起的大国，想要发展得更强大就得多方位了解国际局势和各国态度，从而针对性地制定外交政策。对国民而言，身处全球化的浪潮中，人们想要第一时间了解国际重大事件，各国的态度和举措，进而形成个人世界观以及了解事件对自己生活的影响。通过关键词“跨语言新闻事件”查阅论文得到的结果并不多，这说明该领域的研究还不充分，需要进一步研究。现有材料关注点集中在单一语言新闻报道舆情分析或者某特定行业跨语言新闻事件情感演化分析，因此选择“新闻事件在跨语言传播中情感演化问题研究”作为选题。

理论上，目前针对单一语言新闻事件情感分析和特定行业跨语言新闻事件情感分析的研究比较多，而研究新闻事件在跨语言传播中情感演化问题可以在一定程度上弥补该领域学术的空白，对后续发展有着积极作用。实际中，该研究结果可以指导国家制定外交政策和方针。比如中方媒体从国际关怀和同胞情谊角度报道援助某国的新闻事件，但是受助国媒体可能从自身利益方面抨击中方的援助行为，认定其为经济侵略。由此可以看出，他国新闻报道情感态度能够指导中国对其援助手段进行思考和改变。其次，了解国际事件传播中情感演化有利于及时对不同情感态度做出反映，从而增强本国国际影响力。

1.2 国内外研究现状

1.2.1 单一语言新闻事件获取及情感分析

国内的新闻事件和舆论分析研究有两种。

第一种为中文特定关键词舆论分析，比如 2019 年 6 月上海颁布垃圾分类条例等有关细则，引发民众热议。网络作为信息传播的一个媒介，一定程度上反映了群众对于垃圾分类这个话题的态度。基于以上背景，国内学者李丹妮和梁嘉利用 python 语言编程采集了 2019 年 6 月至 7 月近一个月有关上海垃圾分类的微博博文，接着引用 Snow-NLP 包中的情感分析函数对于新闻素材进行情感舆论分析，得出整体情感积极，负面情绪并不显著的结论。首先，他们从爬虫抓取的大量博文中随机抽取 200 篇进行人工情感标注，整理出正向情感文本和负向情感文本作为训练集参数输入 Snow-NLP 包中自带的函数中训练模型。程序训练模型利用了机器学习中的贝叶斯模型，首先对于训练集文本进行关键词提取，即先对句子分词，之后根据 TF-IDF 算法，算出某个词的词频与其逆向文件频率的乘积，最终得到乘积较高的词语作为特征词，说明正向文本或者负向文本具有以上特征。输入需要预测情感倾向的文本后，模型对待预测文本进行分词后转化为词向量。词向量的每一列代表一个词语，每一列对应的数值为词频。再引入之前的正向负向特征向量进行朴素贝叶斯概率计算，得出情感倾向分类结果[1]。总的来说，该方法是情感分析运用的较为普遍的方法。

第二种方法是针对某种非中文的语言进行新闻舆论分析。比如基于阿拉伯语的网络舆情分析[2]，其新闻采集方法不同于上一种在微博等社交网站中按照关键词批量爬取，而是在谷歌等搜索引擎中爬取，访问由关键词和搜索排行较高的网站域名拼接而成的所有域名链接，本方法利用搜索引擎直接搜索，因此避免了某些社交网站需要登录账号、输入验证码等复杂环节，使得后续信息筛选和数据清洗更加快捷高效，但是存在着因依赖搜索引擎查询相关关键词导致目标信息不够全面的问题。在分析评论文本过程中，这种方法采用分句分词、停用词移除、主题词提取、词组情感倾向标注、主题分类等方式。在停用词移除方面，由于缺乏针对性的停用词表，其采用的是人工词语去除和筛选热频词。最终利用原文中描述的词语提取和频数统计工具获得了排名前 28 的词

语并进行了后续分析。

以上两种方法相似度较高，区别在于，非主流语种现存的可利用资料较少，需要采用算法迁移策略，例如人工标注数据和编写分析工具等[2]。

国外单一的新闻舆情分析研究，大多针对新闻所在国的当地通用语言。例如对于乌尔都语文本情感分析，该语言是巴基斯坦以及西亚很多国家的官方语言。巴基斯坦的学者提出了一种基于递归卷积神经网络的情感分析方法，他们首先在网站采集五类语料，体育、食品、政治、软件和戏剧。接着，由语言专家手工标注语料中提取的样本的情感倾向，正面、负面和中立。利用 N-gram 算法在样本中找出代表某一分类的最重要的词构成词汇集。其创新之处在于利用递归卷积神经网络，用于克服传统神经网络模型如递归神经网络和卷积神经网络模型的一些局限性。传统递归神经网络模型对文本逐词分析，并在隐含层中对句子的上下文信息进行处理，然而，它只倾向于提取句子的最近的词，整体上可能会影响全文的语义。相较于递归神经网络，卷积神经网络引入了最大池的概念，选取文本中重要的一些词语，但是固定的窗口大小使得该方法费力耗时。而新的方法：递归卷积神经网络中，待分析文本未设置固定的词语识别窗口大小，因此，须获得的某词语的关联词组要依靠前后文章内容进行搜索，相邻词结果均由递归神经网络计算得出。结果证明在他们使用的数据集中，递归卷积神经网络模型准确度比其他两个传统神经网络模型高[3]。

1.2.2 跨语言新闻事件获取及情感分析

跨语言的新闻事件情感研究分析相较于单语言来说，难点主要集中在情感表达如何在多语言环境下进行迁移。目前，机器翻译被广泛应用于词汇的映射和关联，但问题在于其会不可避免地产生语义扭曲，从而导致迁移误差，该问题目前亟待解决。

跨语言情感分析会引入一些其他的模型和方法，例如基于多视图的集成学习方法，协同训练方法[4]。第三种方法是跨语言混合模型，该模型使用大量未标记的双语平行语料，选取能够最大化转换双语平行语料的参数，将参数用于学习和获得未出现在情感词典中的情感单词以改善情感词典中的单词

覆盖范围，增强目标语言句子和源语言句子之间的学习对齐关系，从而完成跨语言情感分析任务。

学者陈强总结国内外跨语言情感分析技术提出了一种系统的方法，即基于句法分析的跨语言情感分析方法，运用一种目前主流的句法分析技术：基于树库的统计句法分析，使用从语法分析中获得的语法树模型来分离情感词汇的标题和组成部分即主谓成分作为单词的特征注释和确定所有句子的统计特征权重的标准之一，并且，使用文法树分析丢弃不会表现出情感倾向的连接词，从而提高算法效率。英文情感分析转化为中文分析需要经过源语言分句、句法分析、情感词匹配和权重值计算、情感分类比值或贝叶斯模型阈值计算、贝叶斯分类器训练和源语言文本翻译成目标语言等步骤[5]。中文文本情感结果通过最后一步由英文翻译得出，因此其情感分析准确度也依赖于机器翻译的准确度。

1.3 面临的关键问题

当前阶段问题集中在两个方面：

第一个是小语种单语言语料和模型欠缺问题。网络上现有的数据格式为词典和文本集，目前中文和英文的情感语料库建立的最为完善，网络资源也最丰富，例如中文词库 HowNet,英文词库 WordNet,中文情感词典 BosonNLP 和台湾大学情感极性词典，英文 GI 评价词词典，中英文 HowNet 评价词词典等等。然而对于其他语言比如德语法语意大利语，暂时没有开源的数据库，大部分研究都在实验室中进行。针对不同语言中的不同类型数据：词典和文本集，如何建立情感分析模型，将不同语言的研究统一起来，找到能够跨越语言差异的分析方法，实现跨语言研究是目前面临的第一个问题。

第二个是跨语言文本情感迁移问题，机器翻译准确性不够高，从而影响多语言之间的相互文本转换和情感迁移。

接下来的研究将把以上难点作为重点，提出优化方案，并给出实验验证。

1.4 论文研究内容

研究跨语言新闻事件的获取以及情感演化分析，并实验验证其中的一些方法。新闻事件的获取目前有两种方案，一是大规模批量网页抓取，二是关键词查询指定网站获取。情感分析方法涉及众多模型算法，下文会详细分析不同的模型并比较分析其优劣性。最后的实验部分将采用目前最可行最优的方法对马来西亚语和泰语的新闻事件进行批量抓取和情感演化分析，作为案例提出一种整合并优化现阶段研究的跨语言情感演化分析方案。

1.5 论文组织

第二章介绍跨语言新闻网站上新闻事件的获取方法和情感分析方法，包括中英文新闻事件获取，基于语料库提取关键词，主题情感信息挖掘，情感倾向分析和情感演化分析。第三章提出一种针对跨语言新闻事件进行情感演化分析的方案，包括数据预处理，基于词典的情感分类模型构建，主题词提取和逻辑回归算法结合的情感分析模型构建，第四章将介绍实验过程，即如何获取马来语和泰语两个语种的大量新闻事件按照字段存储在本地的 excel 工作表和 MySQL 数据库中，接着利用第三章提出的情感分析模型进行不同语言文本情感分析，最终通过关键词查询得到按照时间顺序排列的两个语言新闻文本相关信息以及情感分析结果，进而实现基于时间轴的情感分析即情感演化分析。

第 2 章　跨语言新闻事件情感演化分析相关工作

2.1 中英新闻网站上新闻事件获取方法

网络上获取新闻的渠道广泛且丰富，但是人工获取批量数据时存在下载速度慢、浏览耗时等问题。基于上述背景，网络爬虫技术诞生。该技术是一个可以自动下载网页内容的程序，根据特定需要单独收集数据，从中抓取相关网页内容和某些信息，通过访问指定的页面和链接为用户提供资源[6]。总的来说，爬虫的一般流程为通过用户提供的 url 访问网页，获取网页源码后提取所需数据资源，最后将数据格式化存储为文件，便于用户后续进行数据清洗和分析。

Python 被广泛用作实现爬虫的载体因为它含有丰富的开源网络信息抓取模块，语法较为简洁并且拥有爬虫框架 Scrapy。目前 python 爬虫程序分为以下四类，同步、并发多线程、异步和 Scrapy 框架爬取。同步爬取方式即为上文所提到的方法，比如利用 python 的 requests 库中函数访问网站，访问成功后针对网页编码格式使用对应的解析工具。并发爬取采用多线程爬取，例如运用 threading 库，根据实际情况设置并发线程数目。异步与同步类似，不加以赘述。最后一种，利用 Scrapy 框架编写爬虫程序时需要在指定名称文件中填入爬虫代码，实现结构化开发[7]。

国外学者提出了智能爬虫的概念，与传统爬虫相比，它从网页内容中重点提取七个部分，分别是菜单、链接、主文本、标题、概要、必要信息和附加信息。目前基于 Dom 的分割、基于位置的分割和基于视觉的分割是实现自动提取技术[8]的三种方法。比如访问搜狗新闻网站，主页面上获取到很多附属页面链接。接着访问附属链接，如果获得新闻文本不为空则按照新闻文本格式记录提取标题、发布时间和内容等信息，为空则为新的主页面继续访问。实际上该过程类似主新闻网页中包含体育新闻链接，单栏目链接下包含单篇新闻链接，单篇新闻链接包含所需文本信息，运用到了二叉树遍历算法[9]的思想。

2.2 基于语料库的提取关键词方法

目前涉及主题词提取的算法有 LDA 模型和 TF-IDF 模型。LDA 模型[10]认为一个文档含有多个主题，并服从多项式分布，而每一个主题下含有很多词汇并且服从多项式分布。LDA 模型的作用就是根据文章含有的词汇推算出其所属的主题，从而完成主题提取任务。LDA 模型中已观测的变量是那些文档中的词，隐藏变量是话题模型。由此，从文档来推断隐藏话题的问题变为计算后验分布的问题即计算给定文档隐藏变量的条件分布。

TF-IDF 模型[11]实际上计算的是某个词语在文档集中的重要程度。处理过程为，首先对于文档集中所有文章进行分词，分词的标准可以是按照空格划分或者按照标点划分等，同时利用 N-gram 设定分词后每一个分段词语的个数，比如“我喜欢吃苹果”这句话按照 1-gram 划分结果为“我”“喜”“欢”“吃”“苹”“果”，按照 2-gram 划分就会出现“喜欢”“苹果”，划分规则即 n 的选择过程中，不能选的太大，这样得到的词语是稀有词语或是拼写错误词语，无法体现文章主题，但是 n 值也不能选取的太小，太小会得到的停用词太多同样没有效果，因此采用 1-gram 和 2-gram 较为合适。接着以上述过程获得的特征词为列标签，每一行为一篇文章构成矩阵向量，值则按照 TF-IDF 算法计算得出。其计算方法为词语在文章中出现频率即该词在文章中出现次数除以文章总词数文与逆向文件频率就是总文件数目除以包含该词文件数目取对数的乘积。

在上述两种算法基础上，学者王瑞提出了基于 Labeled-LDA 模型的文本特征提取算法[12]，其核心思想是将 LDA 算法提取出来的隐含主题与 TF-IDF 算出的主题词语进行相似度计算，得到相似度较高的词语作为主题词。从而综合两种算法，提高主题词提取的准确性。

国外学者提出了一种与 LDA 算法类似的 EFTM 实体方面主题模型[13]，该模型基于以下三个假设，第一：假设实体具有可以被人工定义的方面特性，第二：多方面实体模型能够提取出文章主题特征，第三：一个实体能够被少量的方面词表示。该模型提取文章主题过程为利用预构建的源实体集，每个实体含有其各方面的属性，不同的属性属于不同的主题，而主题最终对应文章中可能是词语或者短语的实体。

2.3 主题情感信息挖掘方法

主题提取技术是文本挖掘中的一种新技术，用于自动标记和分离单词，短语或句子。网络用户往往针对特定对象发布评论，其情感倾向也常常围绕某个主题。因此基于主题的情感倾向挖掘能够建立主题与情感的关系从而有效地预测文本情感倾向，而传统方法只能判断整个句子或整个章节的情感倾向趋势，而无法分析其深层意义[15]。

2.4 情感倾向分析研究方法

目前主流的情感倾向分析方法有 5 种，分别是基于词典的方法、基于机器学习的方法、基于词典+机器学习的方法、基于弱标注的方法、基于深度学习的方法。

基于词典的方法利用正向负向情感词典对文本进行情感分析预测。其原理为对于待预测文本进行分句分词，遍历词语在情感词典中检索，如果存在则赋值根据规定的权重进行计算从而得出整个句子的情感倾向值，最后通过平均值或者其他方式计算出整篇文本的情感倾向值。

基于机器学习的方法[16]包括数据获取、预处理、建模、分析、可视化等过程，概括为聚类。例如对于电影战狼影评的案例分析，采用 python 进行文本聚类。首先将数据利用 numpy 包转化为数组格式，去除文本中停用词并进行分词和清洗，接着将文本转化为 TF-IDF 特征向量表示。利用 TF-IDF 特征提取得到的向量训练聚类模型，即用 sklearn 的 KMeans 对上述特征矩阵进行训练，分为 10 类。最后对结果进行分析，如簇中正向词语展现出评论情感倾向为正向，说明好评度较高。

基于词典+机器学习的方法[17]是上述两种方法的综合，概括为分类。通过预先人工标注情感倾向的训练集建立分类模型，利用确定集评估模型并根据结果调整参数，最后预测测试集得到情感倾向标注。

基于弱标注的方法即通过文本中体现情感倾向的词标注该文本情感倾向，例如豆瓣影评中的星值，4-5 星为好评，0-2 星为差评。该方法一定程度上减

轻了人工标注数据集的工程量，却忽略了对于词句深层含义的理解。

基于深度学习的方法弥补了弱标注方法的不足，例如基于神经网络的情感分析模型卷积网络分类模型[18-19]、长短期记忆分类模型[20]和深度信念网络分类模型[21]等。

2.5 情感演化分析

广义上的情感演化分析是按照时间发展顺序进行的情感倾向分析，在一定时间段内反映用户对于某一特定主题的情感周期性变化趋势[[22]。目前国内研究较多的是微博情感演化问题，第一步利用第二章网络爬虫技术，第二步利用情感分析工具，例如用 TF-IDF 和 KMeans 聚类方法对于情感词典进行主题词汇和情感词汇提取[23]。在主题词表和情感词表的基础上按照时间顺序对待预测文本进行情感分析和排列从而得到情感演化分析结果。

针对跨语言新闻事件的情感演化分析则较为具体，需要对于含有某关键词的新闻内容、传播新闻国家、新闻发布情感倾向等内容进行分析。例如对于冠状病毒这个关键词相关的报道，前期较为消极，因为病毒感染人数逐渐增加，而后期呈现积极趋势，因为全国合力抗疫使得疫情在一定程度上得到控制。在其他国家，他们对中国进行物资援助，该类事件报道为积极报道，而后来由于疫情在世界范围内爆发，他们可能指责中国，新闻报道角度则为消极。情感演化分析需要按照时间顺序构建新闻情感变化折线图，分析某关键词新闻事件在不同国家的传播规律，挖掘情感倾向改变节点等信息。

第 3 章　跨语言新闻事件情感演化分析模型

3.1 数据预处理

首先，获取情感演化分析所需要的数据由爬虫模块完成。爬虫模块由异步式爬虫程序组成。一个爬虫程序针对一个指定的新闻网站上发布的新闻进行爬取。爬取给定新闻网站的流程简化为定向访问新闻网站首页、获取附属页面链接、访问单篇新闻链接和单篇新闻内容提取保存。爬虫程序中网页访问部分根据网站的构成形式采用 requests 请求方式或者 selenium 框架，数据提取部分均利用 xpath 属性路径进行文本和超链接选取，并保证后续参数修改最小化和自动连续爬取数据。其次，对于每天获取的数据进行数据清洗，包括去除当天重复新闻报道和删除文本内容为空的新闻，最后得到按日期归类的新闻数据，每一条新闻均包含链接、标题、时间和文本四个字段。该框架复用性和可移植性高，对不同语言新闻的数据进行抓取时，只需修改目标网址和对应的 xpath 路径，就能得到相同结构的数据，便于后续对不同语言数据进行整合，因此适用于跨语言新闻数据的采集。

3.2 基于词典的情感分类器模型

对于能够通过词典展开对该语言文本情感倾向分析的情况下，提出基于词典的情感分类器模型。在选取合适的情感分析模型的前期，从 github 上面搜集和整理了泰国和马来西亚新闻的情感词典，并添加上本地新闻数据中的特征词，文件为 json 格式，键为词语，值为权重。对于基于词典的情感分析方法进行了研究，按照原理根据词典中的词语，例如马来西亚词语“salah”，中文意思为“错误的”，权重为-1，马来西亚词语“jenayah”，中文意思为“惩罚”，权重为-1 等，在待预测情感倾向的分好词的文本中进行词语检索，对每一个查询到的词语的权重值进行加和计算，最后对于结果大于零为正向，小于零为负向的方法进行了情感标注尝试。将标注后的新闻与人工标注的新闻进行比较，发现标注准确率较低，准确率约为百分之五十，分析后发现问题在于该方法没

有结合上下文对句子语义进行理解，只是机械地匹配词语，例如“冠状病毒死亡人数增加，但总体得到控制”这句话，“死亡”和“增加”两个词语使句子总体情感倾向计算结果为消极，而根据后半句的语义可知该句子总体情感倾向为积极，这说明基于词典计算的会大大增加情感分析的不准确性。

基于上述方法基础上进行改进，利用词典构建词向量进行分类而不是标注权重进行计算，也就是说，基于词典的情感分类模型利用的是文本分类算法，为搭建分类器而输入的内容是人工构建的分类标签和该标签下的词汇。分类器训练完成后对于输入的待分类文本，首先将文本转换为词向量，接着利用多项逻辑回归算法进行计算拟合，得到分类结果。

搭建分类器时，需要输入前期收集和整理的正向和负向的情感词汇。情感词典的构建方法为，首先在 github 上现有的语言库中下载情感词典，然后提取本地新闻数据的特征词，接着将特征词添加到已有的正向或者负向情感词典中。在跨语言新闻事件情感分析中，需要分别收集和构建多种语言的情感词库，然后聚类分析本地不同语言的新闻文本，找出特征词中共有的词语并添加到情感词典中。例如对于本地马来西亚和泰国部分新闻数据进行聚类，发现共有词“奥林匹克”，马来西亚语为“olimpik”,泰语为“โอลิมปิค”，于是分别在马来语和泰语词典中添加该词。停用词典的构建方法为，首先在 github 上的语言库中下载已有的新闻停用词词典，例如泰语的停用词词典下载路径为：https://github.com/6/stopwords-json，再对已经抓取和保存到本地的新闻数据进行聚类，得到所有的特征词，然后在特征词中人工筛选出一些没有实际含义的词语，例如“他”、“她”等，添加到已有的停用词典中，从而构建出适用于本地新闻数据的停用词词典。利用分类器进行分类时，首先对于输入的数据进行预处理，进行分词、去除停用词和获取特征值，接着将情感词典词汇添加到特征词中，得到词向量。最后，利用词向量和多层逻辑回归模型计算出文本属于正向和负向的概率，概率最大的则是该文本分类结果。

基于词典的情感分类模型的适用场景为能够获取得到人工构建的情感类别词典例如积极类别下含有词语“支持”、“团结”等，消极类别下含有词语“反对”、“谋杀”等。该方法能在不同语言新闻数据中进行迁移，并且由于情感分

析算法相同，结果较为统一。针对不同的语言，词典构建工作需要特定的语言专家对于大量新闻文章的分词结果进行筛选和同义词的添加等步骤，也可以利用网络上官方发布的来源可靠的词典。将一部分马来西亚语和泰语共有词添加到两国的情感词典中后，运用该方法进行马来西亚新闻的情感分类准确率达到了百分之八十五，泰国情感分类准确率为百分之八十。准确率是对十个文档中近一千条新闻利用模型情感分析结果与人工情感分析结果相同的概率，默认人工标注的情感倾向准确。

3.3 主题词提取和逻辑回归算法结合的情感分析模型

对于能够获取到大量文本数据的语言，该方法将文本主题词与情感倾向结合。研究前期采用词袋方式获取主题词导致得到的主题词不纯，例如含有词频过大的停用词和词频过小的稀有词和拼写错误词，对于后续情感分析产生影响。于是主题词提取方法改进为 TF-IDF 算法，首先去除停用词，接着设置下限去除稀有词和拼写错误词，对于词频和文章逆序频率统计结果均采用对数计算，规范化计算结果。

待情感预测文本的候选词是训练集主题词提取得到的，训练集文本格式为一篇文章和人工标注的情感倾向，标注过程由非语言专家的预标注员和语言专家审核人员完成，针对新闻报道的角度进行正向和负向标注，保证训练集数据的准确性和一致性。

训练集，验证集和测试集文本的词向量表示均由上述主题词和 TF-IDF 算法获得，使得结果统一，矩阵维度的一致性较好。利用训练集词向量输入逻辑回归模型进行训练，假设因变量服从伯努利分布，当自变量的值距 0 较远时，因变量的值快速趋近于 0 或者 1，而 0 和 1 可理解为二分类结果，因此该函数可以较好的拟合情感极性分析过程，也是采用逻辑回归模型的原因。利用向量表示之后的训练集训练逻辑回归函数，调整函数中的参数以满足实际情况，就能较好地预测测试集文本向量情感倾向。该方法既包含主题词提取，又包含逻辑回归，从而能够挖掘出主题和情感的联系，改善了传统主题-情感提取中人工查找的费时费力和拟合过程复杂的缺点。

3.4 情感演化分析

在利用上述模型得到文本情感倾向后，按照时间顺序对于某关键词下的所有新闻文本进行排列，建立横向为时间发展顺序，纵向为国家和发布的新闻的情感倾向的坐标轴。其中纵坐标的数值计算结果为积极和消极情感倾向频率的差值，0 到 1 为正向，-1 到 0 为负向，绝对值越大表示情感越强烈。针对不同线性进行分析，就能得到某关键词新闻报道的传播特点。

关键词的选取采用聚类算法，对每日新闻数据进行五个主题和每个主题下五个重点词汇的选取，研究过程中采用了 LDA 算法和 K-Means 方法，对比后确定使用 K-Means 聚类方法，原因是聚类得到的词汇长度较为整齐，不同主题之间的词语差异较大，重点词语代表性较强。聚类的过程为获取指定文件文本，文本 TF-IDF 词向量转换，聚类，每一类下重点词语提取。对于出现次数较多的词语进行统计得到待研究的关键词。聚类结果也用于后续情感演化曲线上转折时间点分析，例如研究导致转折出现的事件等。

提出的情感演化分析方法还需根据实际情况进行调整，例如可视化曲线中横坐标的取值，当应用于时间跨度大的新闻事件中时需要时间分片，而在时间跨度小情况下不分片，按天为时间单位，从而提高情感演化分析的细粒度。纵坐标的计算方式也不是唯一的，可以按照需要分析的角度调整，例如不采用频率而是对于含有关键词的报道进行基于词典的计算：求和或取平均值等。后续研究在目前提出的分析方法框架下适当调整参数以适应不同语言新闻分析的需求是很必要的，尤其是对于情感分析模型的选取，以上提出的两种方法适用于不同的情况，一个基于情感词典，一个基于训练集，实践时需要考虑语言特性，人力开销等方面从而进行选择，结果不一定是最优的，但是可能是最适合研究和具有研究意义和价值的。

第 4 章　实践操作

4.1 实验环境搭建

所有代码均利用 Python3.7 编写，实验环境如下：

操作系统：Windows10

Python 环境：PyCharm Community 2019

Python 包：requests, selenium, lxml, fake-useragent, openpyxl, pandas, datetime, time, fasttext, pythainlp, tqdm, numpy, sklearn, tensorflow 等。

网络环境：无线网和代理软件

数据库：5.7.29 MySQL Community Server

4.2 马来西亚和泰国新闻获取

马来新闻网站有三个，第一个为 https://www.bharian.com.my/拼接上四个属性 berita、sukan、dunia、hiburan 得到:

https://www.bharian.com.my/berita

https://www.bharian.com.my/sukan

https://www.bharian.com.my/dunia

https://www.bharian.com.my/hiburan 四个网站。

第二个马来西亚网站为 https://www.malaysiakini.com/my/latest 拼接上 news、columns、letters、hiburan、sukan 五个属性即得到：

https://www.malaysiakini.com/my/latest/news

https://www.malaysiakini.com/my/latest/columns

https://www.malaysiakini.com/my/latest/letters

https://www.malaysiakini.com/my/latest/hiburan

https://www.malaysiakini.com/my/latest/sukan 五个网站网址。

第三个网站为 https://www.bernama.com/bm，按照属性和页码访问。

爬取过程为：首先访问首页获取该页面上所有单篇新闻链接，如果存在多页，则获取该网站最大页页码，再按顺序从第一页访问到最后一页。

将利用 XPath 获取到的单篇新闻链接存储为列表，之后遍历列表从而访问单篇新闻链接，继续利用 XPath 提取标题、发布时间、文章文本，最后按照链接、标题、发布时间、文本内容四列的格式写入名称包含抓取日期的 excel 表格中。

虽然初期我利用 requests 包来实现爬虫，但是遇到了无法抓取动态页面、修改 cookies 操作繁琐和服务器拒绝访问等问题。因此最后对于上述网站内容爬取均采用了 selenium 框架，这样就无需人工添加 headers 和 cookies，同时还降低了网站反爬虫策略的影响以及被封号风险。从前期使用 requests 屡屡碰壁到用 selenium 实现抓取，我个人比较这两种技术得出的结论是：有时候前端页面展示出来的东西，并不在后端源代码中，无法通过使用 requests 请求获得源码进行爬取，这时候就可以使用 selenium 进行数据抓取，因为其是用真实的浏览器去访问页面，所以出现的内容和我们在前端看到的一模一样。

接着利用相同的框架编写代码对于泰国的四个新闻网站：

https://www.dailynews.co.th

http://www.thairath.co.th

https://www.matichon.co.th 和 https://siamrath.co.th 指定属性和页面进行爬取，并按照格式保存单篇新闻指定内容。

4.3 情感倾向性计算

目前除了中文和英文，其他的语言库并不丰富，能够获取到的数据格式可以分为词典和网页源文本数据集两种。因此针对两种不同情况提出了两种情感分析方法，一种适用于没有人工预标注数据集，只有情感词典的基于分类算

法的方案，另一种适用于有人工预标注数据集，没有情感词典的基于回归算法的方案。为了证明两种模型均适用于跨语言新闻情感分析，试验中将两种方案应用在马来西亚和泰国的新闻数据中，并与其他方法进行比较，说明提出的方案更适合研究本地新闻数据。

试验初期对于马来西亚新闻文本采用机器翻译和调用 snownlp 库函数进行中文情感分析方法。首先利用谷歌翻译工具将人工预标注情感倾向的马来西亚新闻翻译成中文，然后直接调用 snownlp 库中 snowNLP 函数进行情感倾向预测，并将结果写入文件中，最后计算模型预测结果的准确率为 60%，即预测正确的新闻条数占总预测新闻条数的百分比为百分之六十。说明在现有的数据集中，利用机器翻译方法进行马来西亚新闻情感倾向计算准确性不高，不适用于跨语言新闻情感倾向性计算。机器翻译得到的中文新闻内容与原文差异较大，存在语句不通顺，语义错误等问题，导致模型准确率较低。因此不采用机器翻译方法，而是直接在源语言环境下利用最适合该语言的模型进行情感分析，最后对于多语言数据进行整合。

对于泰国新闻文本，研究初期采用朴素贝叶斯中的 Bernoulli 分类方法进行情感预测。首先人工构建训练集和验证集，对于五个文件中的新闻数据进行人工标注情感倾向。然后将标注后的数据集分为训练集和验证集，接着利用训练集训练模型，并适当修改模型参数。最后利用模型预测验证集，计算出模型准确率为 79%，不到 80%。

试验后期对于马来西亚新闻文本情感分析采用 fasttext 文本分类模型，有监督学习方法根据输入标签文本训练模型，最后利用模型预测给定文本所属类别。输入的训练文本一行的格式为__label__标签名，后面跟上属于该类的词语或者句子，作为该标签的特征。情感倾向标签为 pos 和 neg，一部分特征词语为网站 https://github.com/huseinzol05/Malaya-Dataset/tree/master/lexicon 提供的 sentiment.json 中词语，该词语库由马来西亚自然语言处理学者 Husein Zolkepli 创建和维护，专门针对马来西亚情感分析。另一部分特征词语为本地马来西亚语新闻和泰语新闻的部分共有词。利用该模型对于大量 excel 中数据进行情感分析标注 pos 或 neg，之后与人工标注结果进行比对，准确率达到 80%，高于前期针对马来西亚新闻建立的模型的准确率。接着利用相同的方法

对于泰国新闻数据进行情感分类。泰国的情感词典中大部分特征词汇是两国的共有词，该模型准确率为 80%。

利用主题词提取和逻辑回归算法，对于泰语和马来西亚语新闻文本情感分析比上述过程复杂。利用人工标注的训练集和验证集训练模型并计算模型准确率，最后运用模型预测待情感标注文件，该模型经过调试后准确率达到了 88%。该模型与前期试验的朴素贝叶斯模型的训练集和验证集完全相同，词向量提取方法也相同，但是准确率提高了 8%左右，说明该模型更适合分析本地泰语新闻数据。利用该模型对本地马来西亚数据进行情感分析准确率为 85%。以下为操作步骤。

（1）获取训练集、验证集和测试集并转换为 dataframe 格式

从 excel 文件集中读取指定文件，按行处理保存为 all_df 和 test_df 全局变量。

（2）文本转换为 TF-IDF 矩阵向量

利用 vectorize 函数对于输入数据进行向量化。利用 pythainlp.process_thai 提取泰语 token，利用 malaya.preprocessing.SocialTokenizer().tokenize 提取马来西亚语 token，选择 1-grams 和 2-grams,和至少出现 20 次的词语。

（3）计数和数据准备

对于 TF-IDF 处理后的词语计数，并与文本合并。

（4）训练模型

利用模型 liblinear,其他参数参照手册进行调试，选取满足应用条件并且准确率较高的，网站链接为：

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

（5）预测情感倾向

对于待遇测 excel 数据进行预测，将在原 excel 表中增加一列并存入每一行文本的情感倾向预测值。

4.4 情感演化趋势分析

利用 openpyxl 将已经标注了情感倾向的 excel 数据表批量注入 mysql 数据库中。利用 SQL 语句进行查询即可得到某关键词相关按照时间排序的所有新闻数据。我对于新闻条数较多的三个主题冠状病毒、中国和旅游进行了细致的数据分析，分析步骤如下。

利用 navicat 访问 mysql 数据库后，查找含有关键词的所有新闻数据。

接着将关键词输入 navicat 查询选项中，得到按照时间排序的相关新闻，其中 news 后面的表标号为我给定的网站序号 news1、news2 和 news5 为马来西亚新闻，3、4、6、7、8 为泰国新闻。对于数据进行统计，分别得到马来西亚和泰国从 02 月 29 日到 04 月 29 日每日新闻中含有主题词的条数，其中正面情感新闻条数和总体情感倾向值。总体情感倾向值为正向情感概率减去负向情感概率，结果在-1 和 1 之间，负值为负向情感，正值为正向情感，绝对值越大说明情感态度越强烈。最后，我将三个关键词的统计数据转化为折线图展示，并对于特殊点利用聚类等分析方法给出解释，下图中横坐标为日期，纵坐标为情感倾向，蓝色线为马来西亚，红色线为泰国。

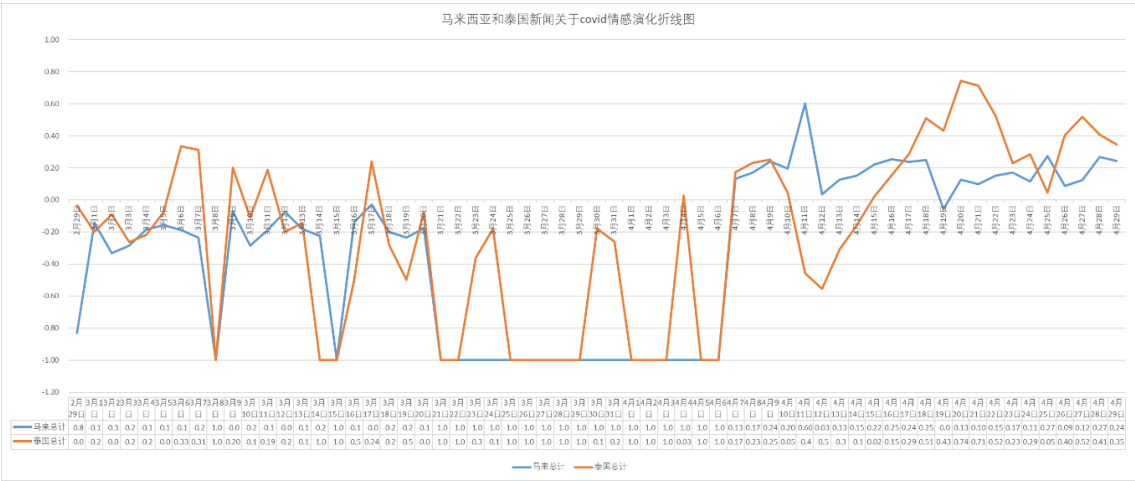

图 1　马来西亚和泰国新闻关于 covid 情感演化折线图

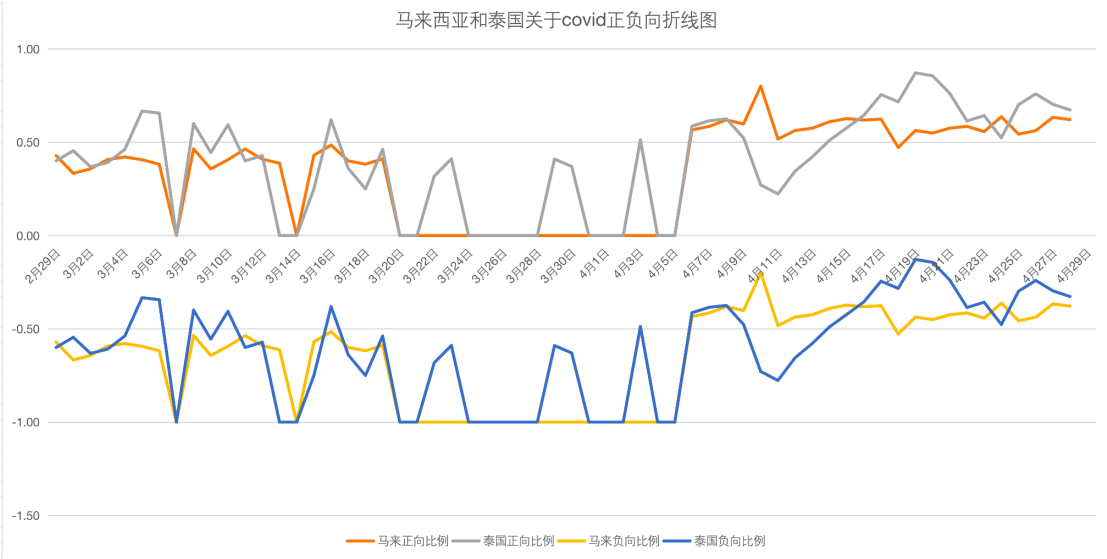

图 2　马来西亚和泰国新闻关于 covid 正负向情感演化折线图

总体观察两个曲线走向发现马来西亚和泰国的新闻报道均从对疫情状态消极的态度转变为较为积极，这与中国最初报道冠状病毒到病毒快速传播和最终得到控制，逐渐好转的趋势相一致，这里主要认为两国的冠状病毒新闻均围绕中国展开，因为对每日新闻进行文本聚类后发现，关键词 covid 与 china 被放到同一个类别下。

由折线图可知，在 04 月 07 日附近两个曲线均从消极转变为积极，发生了明显转变。分析该日期附近新闻得到 03 月 20 日聚类结果中一类为'olimpik', 'tokyo', 'arab', 'saudi', 'covid'，该时间点东京提出推迟 2020 年奥运会的方案，由于国外冠状病毒形势逐渐严峻。而 04 月 07 日与 08 日聚类结果中均有一类为‘seniman', 'malaysia', 'rm', 'covid', 'seni’，新闻报道集中在马来西亚本国疫情情况上，并且由于处于初期，相关政府也因为较早关注了中国局势，因此较早对本国采取了控制措施，所以疫情发展并不严重。泰国分析结果与马来西亚类似，因此两个曲线走向较为一致。

接着对于关键词中国的马来西亚和泰国数据进行统计和可视化，得到的折线图如下。

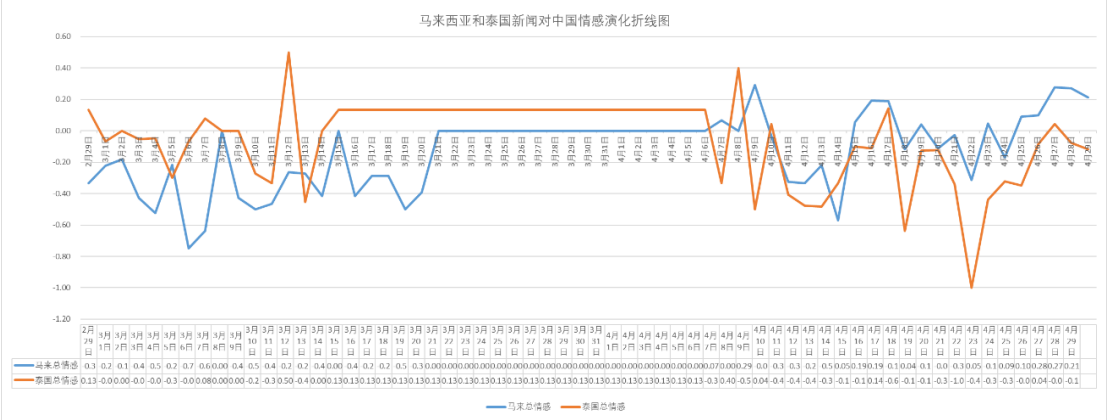

图 3　马来西亚和泰国关于中国新闻演化折线图

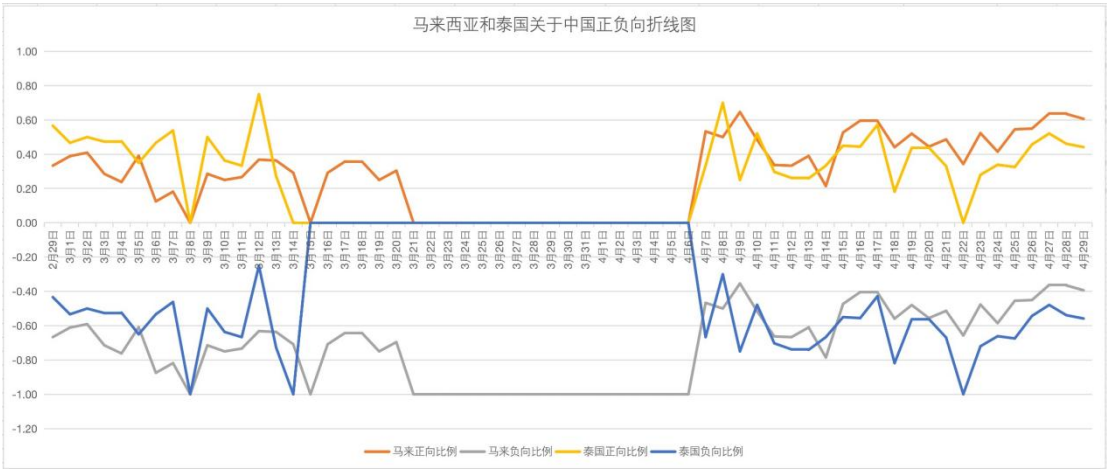

图 4　马来西亚和泰国关于中国新闻正负向演化折线图

总体上观察可以发现马来西亚的波动与泰国异步，并且延迟几天，但总体的趋势仍相同，情感倾向程度不大，大部分在-0.6 与 0.4 之间上下浮动。不过泰国新闻在 03 月 12 日和 04 月 08 日达到波峰，04 月 22 日达到波谷，分别对以上三天含有关键词的新闻进行聚类，03 月 12 日含有冠状病毒的一类中包含关键词控制，具体新闻报道内容为专家预测疫情会得到控制，另一类包含市场和领先，相关新闻报道中国市场回暖趋势等积极内容，因此情感倾向为积极，并且程度相对较高。04 月 22 日达到波谷，聚类结果中有能力不足和异常等词汇，相关报道为冠状病毒在西半球蔓延严重，特别是欧洲国家和美国，泰国也处于案例上升的时期，并且报道包含中国这个关键词，因此情感倾向比较消极，程度达到了波谷。

最后一个分析关键词为旅游，马来西亚和泰国的旅游业非常发达，新闻中

也占有较大篇幅，因此作为研究对象，挖掘疫情下两国旅游业的情况，统计数据后得到的折线图如下。

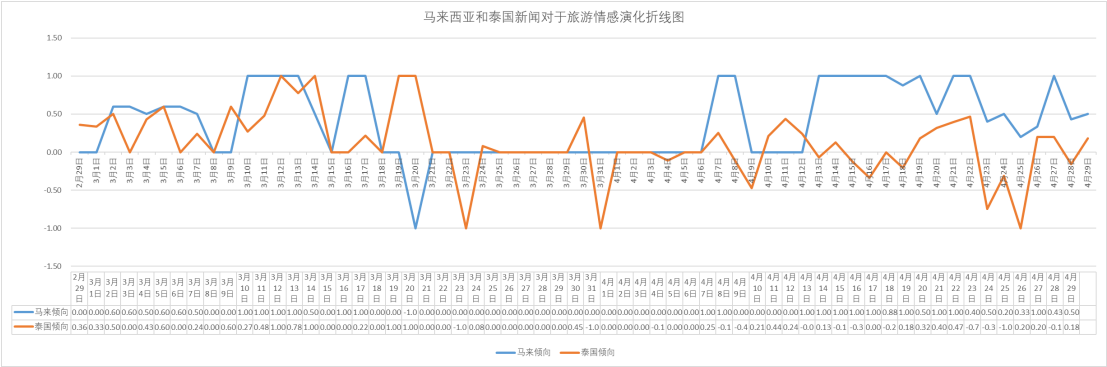

图 5　马来西亚和泰国对于旅游的情感演化折线图

总体而言，马来西亚和泰国新闻报道对于旅游业大部分处于积极态度。其中马来西亚情感波谷出现在 03 月 20 日，泰国出现在 03 月 23 日、03 月 31 日和 04 月 25 日。根据聚类结果，03 月 20 日马来西亚关于旅游的报道包含冠状病毒，具体新闻为疫情下尽量不要出行，所以对于旅游的报道情感呈现负面，不支持旅游。03 月 23 日泰国旅游报道涉及旅游业在疫情下损失惨重等内容呈现负面情绪。03 月 31 日为关闭普吉岛机场和封锁一些旅游景区的新闻，对于旅游也呈现消极态度。04 月 25 日泰国政府一些部门提出开放一些地区的旅游景区，方案最终取消因为考虑到定期评估和开放区域大小等问题，目前还未做好准备。

第 5 章　总结和展望

5.1 总结

前三章理论介绍和第四章实践操作给出了跨语言新闻事件情感演化问题的解法，包括完整的数据获取，数据清理，属性提取，数据分析和处理等方面。对于涉及的算法进行了解释和实践，具有较好的综合性和前瞻性。

5.2 展望

未来工作将集中在建立更准确的情感分析模型和更高效的获取新闻内容两大方面上。争取继续抓取数据添加到数据库中，扩大数据库规模和优化查询结果。

参考文献

[1] 李丹妮，梁嘉. 新浪微博中的“上海垃圾分类” 议题文本挖掘研究 ——基于 Python Snow NLP 的舆情分析[J].东南传播. 2019，9：93-95

[2] 李振华. 阿拉伯网络舆情分析[J]. 阿拉伯世界研究，2013, 3:107-120 页

[3] Zainab Mahmood (Data Curation)a , Iqra Safdera , Rao Muhammad Adeel Nawabb , Faisal Bukharic , Raheel Nawazd , Ahmed S. Alfakeeha , Naif Radi Aljohanie , Saeed-Ul Hassana. Deep sentiments in Roman Urdu text using Recurrent Convolutional Neural Network model[J]. Information Processing and Management 57 (2020) 102233

[4] 陈强. 跨语言情感分析方法研究[J]. 中国博士学位论文全文数据库. 2017, 5

[5] 陈强，何炎祥，刘续乐，孙松涛，彭敏，李飞. 基于句法分析的跨语言情感分析[J]. 北京大学学报(自然科学版), 2014, 1:50(1):55-60

[6] 陈清. 基于 Python 的网站爬虫应用研究[J]. 通讯世界. 2020, 1:202-203

[7] 翟普. Python 网络爬虫爬取策略对比分析[J]. 电脑知识与技术. 2020, 16(1):29-34

[8] Erdinç Uzun, Edip Serdar Güner, Yılmaz Kılıçaslan, Tarık Yerlikaya, Hayri Volkan Agun. An Effective and Efficient Web Content Extractor for Optimizing the Crawling Process[J]. Software: Practice and Experience 2014, 44:1181-1199

[9] 尹文航. 二叉树遍历算法与赫夫曼编码的实现过程[J]. 通讯世界. 2019, 4:63-64

[10] WEI Y L，WANG W，WANG B L，et al．A method for topic classification of webpages using LDA-SVM model[C]. 2017 Chinese Intelligent Automation Conference(CIAC2017). 2017:589-596

[11] KIM D, SEO D, CHO S. Multi-co-training for document classification using various document representation: TF-IDF and Doc2Vec[J]. Information Sciences, 2019, 477: 15-29

[12] 王瑞，龙华，邵玉斌，杜庆治. 基于 Labeled-LDA 模型的文本特征提取方法[J]. 电子测量技术. 2020, 43(1)141-146

[13] Chuan Wu, Evangelos Kanoulas, Maarten de Rijkeb. Learning entity-centric document representations using an entity facet topic model[J]. Information Processing and Management, 57 (2020) 102216

[14] 张晔，孙光光，徐洪云，庞婷，曲潇洋. 国外科技网站反爬虫研究及数

据获取策略研究[J]. 竞争情报. 2020, 16(1)24-28

[15] 朱晓霞, 宋嘉欣, 张晓缇. 基于主题挖掘技术的文本情感分析综述. 情报理论与实践[J]. http://kns.cnki.net/kcms/detail/11.1762.G3.20190715.0941.004.html

[16] 饶泓，姬名书，朱剑. 机器学习课程案例设计与分析——以舆情智能分析案例设计为例[J]. 实验技术与管理. 2019, 36(6)152-180 页

[17] JIE J, RUI X. Microblog sentiment classification via combining rule-based and machine learning methods[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2017,53(2):247-254

[18] KIM Y. Convolutional neural networks for sentence classification[J]. Eprint Arxiv, 2014

[19] 张海涛，王丹，徐海玲，等. 基于卷积神经网络的微博舆情情感分类研究[J]. 情报学报，2018，37(7)：695-702

[20] ZHU X, SOBHANI P, GUO H. Long short-term memory over recursive structures[C]. Proc of Int Conf on Machine Learning. New York:ACM, 2015:1604-1612

[21] RUANGKANOKMAS P, ACHALAKUL T, AKKARAJITSAKUL K. Deep belief networks with feature selection for sentiment classification[C]. International Conference on Intelligent Systems. IEEE, 2017: 9-14

[22] 李超雄，黄发良，温肖谦，李璇，元昌安. 基于动态主题情感混合模型的微博主题情感演化分析方法[J]. 计算机应用. 2015,35(10):2905-2910

[23] 朱晓霞，宋嘉欣，孟建芳. 基于动态主题—情感演化模型的网络舆情信息分析[J]. 情报科学. 2019, 37(7)72-78

附录：跨语言新闻事件情感演化问题文献综述

一、前言

目前网络是信息交流的最大载体，也是获取新闻的最优途径之一。运用网络爬虫技术，将不同语言新闻网站上的内容自动批量下载和保存[1]，而后对新闻数据进行情感演化分析，得到在不同国家的新闻传播的趋势和舆论走向[2]。通过研究含有关键词的新闻事件在不同国家被报道的情感倾向变化趋势，能够指导外交政策的制定和控制舆论走向。

围绕跨语言新闻事件情感演化问题，阅读了以下参考文献：

[1] 陈清. 基于 Python 的网站爬虫应用研究[J]. 通讯世界. 2020 年 01 期 202-203 页

[2] 李超雄，黄发良，温肖谦，李璇，元昌安. 基于动态主题情感混合模型的微博主题情感演化分析方法[J]. 计算机应用. 2015,35(10):2905-2910

[3] 洪旭东. 有色行业跨语言新闻事件信息获取与分析方法[D]. 昆明理工大学，2017

[4] 朱晓霞，宋嘉欣，孟建芳. 基于动态主题—情感演化模型的网络舆情信息分析[J]. 情报科学. 2019 年 07 期 72-78 页

[5] 朱晓霞，宋嘉欣，张晓缇. 基于主题挖掘技术的文本情感分析综述. 情报理论与实践. http://kns.cnki.net/kcms/detail/11.1762.G3.20190715.0941.004.html

[6] 李超雄，黄发良，温肖谦，李璇，元昌安. 基于动态主题情感混合模型的微博主题情感演化分析方法[J]. 计算机应用. 2015,35(10):2905-2910

[7] KIM Y. Convolutional neural networks for sentence classification[J]. Eprint Arxiv, 2014

[8] RUANGKANOKMAS P, ACHALAKUL T, AKKARAJITSAKUL K. Deep belief networks with feature selection for sentiment classification[C]. International Conference on Intelligent Systems. IEEE, 2017: 9-14

[9] ZHU X, SOBHANI P, GUO H. Long short-term memory over recursive structures[C]. Proc of Int Conf on Machine Learning. New York:ACM, 2015:1604-1612

二、文献综述

1. 单一语言新闻事件获取

1.1 新闻分类

新闻分类指的是将新闻网站上大量新闻按照预先定义好的主题类别分类，如军事、政治、经济等。利用机器学习方法，将文本进行分词等操作形成特征向量，然后用已经分类好了的训练集调整参数得到模型，再对测试集进行分类，从而实现将海量新闻事件分成不同的类别[3]。爬取到的新闻数据为一个网站上未经分类的所有新闻，利用该方法可以将新闻分类，从而利于后续的热点新闻提取工作。不过实际过程中，更简便的方法为链接解析法，由于获取的新闻数据包含新闻链接，通过解析链接构成能够直接将新闻分类[2]。

1.2 新闻聚类

新闻聚类指的是内容相近的新闻自动分到同一类，与新闻分类的区别是没有预先规定的类别。过程是首先将新闻文本表示成特征向量，再进行相似度计算，相似度高的分为一类，目前聚类方法有 LDA、K-means 等[3]。K-means 方法能够不断迭代优化聚类结果，因此应用更为广泛。聚类方法与特征词提取结合可以用于关键词聚类，得到不同类别下最有代表性的词语，从而指导关键词的选择。

2. 跨语言新闻事件分类

如果将单一语言新闻分类方法运用到跨语言新闻事件分类中，需要针对每种语言构建训练语料和分类模型，缺乏整体性，因此提出跨语言新闻事件文本分类方法，即通过一种语言的新闻语料，一个分类器实现跨语言新闻事件分类[3]。首先，根据语言专家建立的语料库，建立跨语言特征向量，训练得到分类模型，再集中地将跨语言文本统一分类[7]。但是过程中可能利用机器翻译进行跨语言词库建立，因而结果会不准确，并且由于需要专家建立多语言库，将耗费大量人力和时间，不易于短期项目的开展。

3. 基于主题挖掘技术的文本情感分析

目前情感分析可分为静态情感分析和动态情感分析两种，静态情感分析的目标是得到网民对于某一事件的态度倾向，支持还是反对，忽略了情感波动以及外界因素比如第三方政策的影响。而动态情感分析则弥补了静态情感分析的不足，总体上将事件按照时间切片，再在每个时间片内进行情感倾向分析，最终整合到时间轴上，得到时序排列的情感演化过程[4]。

3.1 基于主题的静态情感倾向分析

主要任务是分析出该主题下文本的情感倾向。利用分类器或者情感词典完成对于整篇文本的情感分类或者情感极性计算。具体实现步骤包括关键词提取，即手动标注传达情感态度的词褒义词贬义词程度词等；模式提取和匹配，即通过情感单元的构建提取出情感关键词，例如名词加情感词加名词是一个情感单元，就能提取出中间的情感词；文本情感计算，即按照计算规则和情感词匹配情况计算整篇文本情感倾向[5]。实际应用来说，Snow-NLP 就是一个情感分析的 python 工具包，通过情感词典得到整篇文本的情感倾向值。但是该方法应用范围较为局限，需要后期维护，不断扩充词典和更新标注。其准确率在不同文本中会有较大区别，因此不适合作为固定的情感分析工具。

3.2 基于主题的动态情感演化分析

情感演化分析是在情感倾向分析的基础上更进一步的工作，要将前一步情感倾向分析结果按照文本发布的时间顺序进行整合，在时间轴上看情感状态变化的趋势，从而动态看出主题变化和情感变化的关系，得到深层情感挖掘与分析结果[5]。

该领域目前研究成果为以下三类：第一类模型将情感与主题融合，例如 LDA 主题模型，该模型由以 Blei 为代表的几位学者于 2003 年提出，其核心思想为提取出文档中的主题以及每个主题出现概率，基于其认为文档的编写过程从几个主题中选择关键词汇连成句子段落，LDA 模型则是文档形成的逆过程[8]。主题与情感之间有着密不可分的联系，因此提取出主题就能判断出情感倾向。而由于该模型将文档理解为单个词的集合的不合理性，Lin 和 He 在此

基础上进一步提出了联合情绪主题模型，将文档中情绪和主题一起提取出来。国内其他学者如孙艳和闻斌等人也提出了其他融合主题和情感的模型，但都在情感演化分析上有所不足。第二类是加入时间信息的主题模型，例如 Blei 和 Lafferty 提出的动态主题模型，得到大量文本在时间顺序上主题的变化。Wang 和 Maccalum 提出连续时间演化模型，随着时间顺序，主题的产生和相关性发生变化。这一类只考虑了时间和主题两个因素，在情感分析上有所不足。第三类模型融合三个因素，即时间主题和情感，建立基于主题的情感演化分析模型[9]。例如，黄卫东、林萍和董怡基于 PLSA 模型中主题-特征词-情感词联系抽取得到主题、特征词和情感词，从而进行情感演化分析。李超雄和黄发良在主题情感混合模型上加入情感周期划分，提出动态的主题情感混合模型。情感周期是指情感在第三方作用下出现的周期性变化，具体有情感周周期、情感月周期、情感年周期等，比如情感周周期指人一周七天的情感变化规律[6]。李慧和胡云凤先将文档进行时间段划分，统计每个时间段上情感极性，再运用上述动态主题情感混合模型得到主题、特征词和情感词。而以朱晓霞为代表的学者在其文章中认为以上三类在现实舆情分析中存在不足，因此首次提出基于动态主题-情感演化分析模型，基本思路为先运用语义标注得到情感单元，接着用聚类方法从上一步的情感单元词表中得到主题词，再在划分的每个时间段中统计文本情感强度，从而实现基于主题的情感演化分析[4]。

三、总结

综上所述，针对跨语言情感演化问题的研究虽然较少，但是对该问题下的新闻事件获取和情感分析两个子问题的研究已经有很多成果。所以将已有的技术加以评估选择应用到该问题下，是目前待解决的问题。毕业论文将优化整合前期子问题的研究成果，优化得到分别适用于词典和数据集的两种跨语言情感分析模型，再根据动态主题-情感演化模型[4]的思想，实践操作完成泰国和马来西亚新闻事件情感演化问题的实例分析。

致谢

感谢我的指导老师梁野老师从选题、开题、论文撰写到答辩以来对我的细心指导。每次遇到难题时，梁老师都能第一时间给我专业的解决方案，让我受益匪浅。特别感谢梁老师在大一时教我们 C 语言课程，给我的编程道路打下了基础，以及大二的数据库操作基础课程，让我熟练掌握了一些 SQL 基本操作命令，对于 Mysql 数据库的操作也不再陌生。衷心感谢计算机系的所有老师，谢谢你们的耐心教导和辛勤付出，我非常荣幸能成为你们的学生，并度过非常充实和有意义的四年。希望疫情过后早日回到校园，当面向老师们表示敬意与问候。