硕士专业学位论文

基于机器学习的政协提案和相关舆情的分析

Analysis of CPPCC Proposal and Related Public Opinion Based on Machine Learning

作者：刘一健

导师：赵永祥

北京交通大学

2019 年 5 月

学位论文版权使用授权书

本学位论文作者完全了解北京交通大学有关保留、使用学位论文的规定。特授权北京交通大学可以将学位论文的全部或部分内容编入有关数据库进行检索，提供阅览服务，并采用影印、缩印或扫描等复制手段保存、汇编以供查阅和借阅。同意学校向国家有关部门或机构送交论文的复印件和磁盘。学校可以为存在馆际合作关系的兄弟高校用户提供文献传递服务和交换服务。

（保密的学位论文在解密后适用本授权说明）

学位论文作者签名：刘一健　　导师签名：赵永红

签字日期：2019年5月27日　　签字日期：2019年5月27日

学校代码：10004 密级：公开

北京交通大学

硕士专业学位论文

基于机器学习的政协提案和相关舆情的分析

Analysis of CPPCC Proposal and Related Public Opinion Based on Machine Learning

作者姓名：刘一健 学 号：17125039

导师姓名：赵永祥 职 称：副教授

工程硕士专业领域：电子与通信工程 学位级别：硕士

北京交通大学

2019 年 5 月

致谢

本论文的研究工作是在我的导师赵永祥副教授的悉心指导下完成的。赵永祥老师精益求精的学术精神、豁达的人生理念以及风趣幽默的表达方式和生活态度，深深地感染和激励着我不断进取，对我的工作和学习有着很大的影响，赵老师的言传身教和谆谆教诲也将使我受益终生。在此衷心感谢两年来赵永祥老师对我的悉心指导和关怀。

感谢实验室里的所有老师。衷心感谢郭宇春老师、李纯喜老师、郑宏云老师、张立军老师、陈一帅老师和孙强老师在我的研究生学习阶段对我的指导与关怀，我所有的科研成果都凝结着各位老师的辛勤汗水。特别感谢郭宇春老师、李纯喜老师和郑宏云老师对我的科研项目的分析与指导，在此向各位老师表示诚挚的谢意。

其次，在实验室学习生活和撰写论文期间，张大富师兄、张闯师兄、胡玮师兄、胡安民、陈泽龙、高志朋、赵红娜、苏迪等同学对我的研究工作给予了热心帮助，在此向他们表示我的感谢之意。

另外，感谢国家自然科学基金的支持，在国家自然科学基金项目《基于熵理论的信息匹配网络测量与建模》（基金号：61872031）的支持下，我的研究工作有了更多的思路。

最后，特别感谢一直无微不至的关心、支持我的父母和女朋友，正是他们积极的鼓励和默默的奉献，才使得我顺利完成学业，成为社会的有用之才。

摘要

全国政协提案是我国政治制度非常重要的机制之一，每年全国各级政协委员都要提出提案，仅北京市 2018 年公开的提案就有 798 件，全国各级政协委员提出的提案总数更多。采用技术手段对政协委员形成的提案进行热点主题发现，并根据这些热点主题进行舆情统计分析，可以挖掘相应的社情民意，为政协委员提供技术信息的参考。

目前，关于提案的热点主题发现和采用技术手段对热点主题进行舆情统计的相关研究尚未见到。本文设计了一套政协提案及其相关舆情分析系统，为政协委员提供信息技术支持。本文主要工作包括以下几个方面：

（1）对政协提案划分主题并提取关键词。编写网络爬虫程序，从政协提案网站采集了提案数据；根据政协提案的结构特点对提案进行向量化表示，使用 K-means 聚类算法对提案进行聚类，每一类表示一个主题；设计了两种关键词提取算法从每个主题中分别提取出三个关键词，分别简称“长词”和“短词”，并设计对比实验分析了两组关键词的有效性，结果表明“长词”比“短词”更能反映主题内容。

（2）设计、训练情感分类模型并预测所有未标注数据的标签。开发爬虫程序，采集了每个“长词”的微博舆情数据并保存为结构化文本格式；设计了基于双向 LSTM 的情感分类模型，训练模型，在测试集上达到了 90.45%的准确率，远远高于基于传统机器学习算法的情感分类模型在该数据集上的测试准确率。

（3）对政协提案的相关舆情进行统计并可视化。在上述工作的基础上，对获取的微博舆情数据进行了统计：从关注度演进趋势和关注度大小、情感演进趋势和情感倾向等角度对每个主题的相关舆情进行了统计分析。

关键词：主题发现；关键词提取；爬虫；情感分类；舆情分析

ABSTRACT

The National Committee of the Chinese People's Political Consultative Conference (CPPCC) is one of the most important mechanisms in China's political system. Every year, members of the National Committee of the Chinese People's Political Consultative Conference will submit proposals. There are 798 proposals published by the Beijing People's Political Consultative Conference website in 2018. There is much more proposals submitted by CPPCC members across the country. Using technical method to discover hot topic of proposals and to conduct statistical analysis of public opinion, we can explore the trends related public opinion.These job can provide technical information reference for CPPCC members.

At present, relevant researches on the hot topic discovery of the proposal and on the public opinion statistics of hot topics have not been seen. We design a set of CPPCC proposals and related public opinion analysis systems to provide information technology support for CPPCC members. The main work of our paper includes the following aspects:

(1) We divide the topic and extract keywords of the CPPCC proposal. We realize a web crawler program and fetch the proposal data from the CPPCC proposal website; We vectorizes the CPPCC proposal according to its structural characteristics, and use the K-means clustering algorithm to group the proposals into categories, where each categorie representing a topic; We design two keyword extraction algorithms to extract keywords from each topic, which are referred to as “long words” and “short words” respectively. We design a comparison experiments to analyze the validity of the two sets of keywords. The results show that “long words” are more effective than "short words" to describe the hot topic of proposals.

(2) We design and train a sentiment classification model to predict label for all unlabeled data. We develop a crawler program to fetch weibo lyric data for each "long word". We design a sentiment classification model based on Bi-directional LSTM to predict label for all unlabeled data. This classification model achieves an accuracy of 90.45% on the test set, which is much higher than the accuracy of the sentiment classification model based on the traditional machine learning algorithm on the same test set.

(3) We statistic and visualize the relevant public opinion of the CPPCC proposal. On the basis of the above work, analyze the data of the acquired weibo lyrics and statistic the the related public opinion of each topic, including trend of attentional evolution, the

attention level, the trend of emotional evolution and the sentiment orientation.

KEYWORDS: Topic discovery; Keyword extraction; Web crawler; Sentiment classification; Public opinion analysis.

目录

1 绪论

本章主要介绍本文的研究背景和意义以及国内外研究现状，简单概述本文的研究内容和结构安排。主要分为四部分：第一部分介绍本文研究的背景和意义，第二部分介绍与本文研究内容有关的国内外研究现状，第三部分从总体上介绍本文的主要研究内容，第四部分介绍本文的结构安排，并对每一章的研究内容做简短介绍。

1.1 研究背景和意义

全国政协提案作为我国政治制度的重要机制之一，在我国的各项事业中发挥着重要作用。每年全国各级政协委员都要参加政协会议并提出政协提案，仅北京市2018年公开的提案就有798件[1]，全国各级政协委员提出的提案总数更多。

提案的形成过程需要花费大量的时间和精力。提案的形成是一个系统工程，包括政协委员提出问题，对存在的问题进行深入调研并分析，最后针对问题和分析结果提出切实可行的建议并形成书面文稿。因此全国各级政协委员在提案的形成工作上会花费大量的时间和精力。

在互联网络时代，利用技术手段挖掘提案的热点主题和相关舆情可以为政协委员提供技术信息参考，节约时间和精力。采用技术手段对每年全国各级政协委员形成的大量提案进行热点主题发现，可以为政协委员从整体上把握提案的关注点提供技术信息的参考，从而节约政协委员将来的提案形成时间；政协提案提出后，提案相关话题会在互联网媒体上引起讨论，各个组织、机构和部分网民会在门户网站、新闻媒体、新浪微博、微信公众号等各种媒体上面发布有关提案的消息并且发表相应的看法和评论，利用舆情分析手段对这些评论信息进行分析，可以挖掘相应的社情民意及其变化趋势，从而使得政协委员可以在短时间内快速把握提案的舆情效果，节约政协委员形成新的提案的时间和精力。

目前，关于提案的热点主题发现和采用技术手段对热点主题进行舆情统计的的相关研究尚未见到。对政协提案，目前的研究方向主要包括以下几个方面：研究政协提案在政府决策中的作用[2]；研究政协提案的采纳和办理程度[3]；研究政协提案的信息化管理系统的设计[4]等。这些研究旨在分析政协提案对政府决策的影响、如何提高政协提案的受理、反馈、实施效率等，并没有对政协提案的热点主题和相关舆情进行挖掘分析。

本文拟获取公开的政协提案数据，采用机器学习、深度学习、自然语言处理等相关技术，对政协提案及相关舆情进行挖掘分析，为政协委员提供信息技术支持。具体如下：采用主题划分方法将主题相同的提案划分为同一类，从而发现提案中的热点主题，从每类主题中提取出一组关键词；利用关键词获取与主题相关的微博数据，使用微博数据分析人们对不同主题的关注度大小及演进趋势、情感倾向及演进趋势。本文的工作可以为政协委员提供信息技术支持。

1.2 国内外研究现状

本文的研究工作是获取公开的政协提案数据，对数据进行挖掘分析，将相同主题的提案划分为同一类别，并从划分好的主题类别中提取出每个主题的关键词，然后获取与每个主题相关的微博数据，利用微博数据分析每个主题的提案所引起的舆情变化。因此，本论文有三个相关研究领域：主题发现、关键词提取和舆情分析。

本节主要介绍国内外关于主题发现、关键词提取、舆情分析方面的发展现状。

1.2.1 主题发现研究现状

主题发现也称主题抽取或主题识别，广义的主题发现是指从各种类型的信息源如文本、图片、语音中发现代表性信息的方法；狭义的主题发现的研究对象只是文本数据，专指从文本数据中发现主题的方法[5]。由于本文的研究对象是政协提案数据和微博数据，均为文本数据，因此只介绍关于文本数据的主题发现的研究现状。

在国外，Cheung 等人认为可以使用聚类方法对文本集进行聚类，每一类表示一个主题，然后使用聚类的质心表示每个主题[6]，但是只用聚类质心不能全面的表示主题内容；Perkowitz 等人为了寻找更加有效的聚类方法，探索了一种统计聚类算法，然后评估数据的每种属性对于聚类的重要程度，将权重最大的属性作为聚类结果中每个类别的主题描述[7]，但是只使用权重最大的一个属性描述主题也不全面；Mehrotra 等人认为由于短文本长度较短，包含的信息量非常有限，直接对短文本进行主题发现，一般情况下不能取得很好的效果。因此他们将多个短文本进行组合，从而将短文本扩展成长文本，之后使用主题模型隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)对组成的所有长文本进行主题研究，使用这种方法对短文本进行主题发现可以实现更高的纯度和互信息[8]。

国内，王李东等人通过改进传统的 LDA 主题模型方法提出了 TC_LDA 模型，并将改进后的新模型应用于普通文本语料库和数字图书语料库的主题发现，这种模型通过对图书中的目录和正文进行联合主题建模，实现了对数字图书语料库和

文本语料库的主题发现，使每个主题中的文本类别相同[9]；郭建永等人对传统的聚类算法进行改进，研究出一种增量层次聚类算法，这种算法是一种多层聚类算法，通过生成包含主题和副主题的层次树进行主题划分，使主题相同的文档集和文档摘要在同一个类别中[10]。

1.2.2 关键词提取研究现状

关键词提取也称为关键词抽取或关键词标注，包括单文档关键词提取和多文档关键词提取，是指从单文档或多文档中提取出一组能够反映文档主题的词或词组。

国外很早就开始研究关键词提取技术。在 1957 年，美国 IBM 公司的 Luhn 开始研究文献自动标引方法，提出了一种基于词频统计的方法，这标志着关键词提取技术研究的开始[11]，但单纯的将出现次数最多的词作为关键词没有考虑句子的语义信息，且由于停用词的存在，使用词频法提取出的关键词可能没有意义；Hulth 使用有监督的机器学习方法提取关键词，关键是在算法中加入了合成特征如短语块进行训练，将这种方法提取出的关键词和简单统计词频提取出的关键词与专家标注的关键词对比，结果表明作者的方法效果更好[12]；Fortuna 等在主题本体的构建中提出使用支持向量机的方法对主题分类来选择关键词[13]；Xu 使用聚类算法提取关键词，在聚类过程中考虑关键词的长度、聚类中心的窗口大小等参数，实现了将 F-score 值提高 7.5%的优化效果[14]。

国内的马力等人认为如果将文档中的词语构造成网络图，则其中聚类性强的词语对关键词的提取更重要，因此作者提出了一种度量词语聚类特性变化的变量来测量词语的重要程度，这种方法提高了对词语的重要性判断能力，作者称这种方法为基于小世界网络的关键词提取算法[15]；陈忆群等人设计了一种关键词自动抽取算法，使计算机能够像人类专家一样，利用知识库对目标文本进行学习和理解，从而实现自动抽取关键词，作者的实验结果表明，使用这种方法在公开数据集上的效果很好[16]。

1.2.3 舆情分析研究现状

舆情分析是指针对某一个特定问题，利用计算机技术对这个问题的舆情进行深层次的分析研究，得出人们对这个问题的关注度变化、情感变化、传播路径等，从中分析出该问题产生的社会影响力及其变化趋势。

本文主要从关注度分析和情感分析两个方面研究与政协提案相关的舆情，由于关注度通过对舆情数据进行数量统计进行分析，没有用到具体的技术和算法，因此下面只介绍情感分析方面的研究现状。

情感分析方法主要包括基于情感词典的情感分类、基于传统的机器学习算法的情感分类以及基于深度学习算法的情感分类。

（1）基于情感词典的情感分类

在 2001 年，Huettner 等人将一系列情感词进行正负极性标注并构造成情感词典用于未知数据中情感词的判定[17]；Shanahan 等人利用情感词典找出舆情数据中的情感词并分析词语之间的搭配关系[18]。Esuli 等人则在 WordNet 英语语义词典的基础上，构建了一个迄今为止最著名的英文情感词典 SentiWordNet[19]；在使用情感词典进行中文文本情感分析方面，娄德成与姚天防通过分析文档主题和词语之间的搭配关系计算词语极性[20]。

（2）基于传统机器学习的情感分类

在 2002 年，Pang 等人首次将几种机器学习算法应用于电影评论数据的情感分类，作者使用几种不同的特征选择方法在最大熵模型、朴素贝叶斯模型、支持矢量机模型上进行了情感二分类[21]；在 2005 年 Pang 等人又进一步实现了电影评论的情感三分类和四分类[22]；Whitelaw 等人将文本使用矢量空间模型表示，在特征选择上，使用形容词及其修饰的名词作为特征，在支持矢量机模型（Support Vector Machine, SVM）上的二分类准确率达到 90.2%[23]。李思等人使用条件随机场对文本情感词和情感倾向进行分析[24]；徐军等人以中文新闻评论文本数据为研究对象，选择其中的词频、否定词等作为特征，用于基于朴素贝叶斯模型和最大熵模型的情感分类器，实验取得了不错的效果[25]。

（3）基于深度学习的情感分类

2011 年，Socher 提出了循环神经网络模型（Recurrent Neural Network, RNN），可以将词矢量组合为句矢量，并能够实现对文本中核心词的组合词进行记录和修改，使模型具有学习自然语言运算符的能力[26]；为了解决 RNN 模型存在的长期依赖问题，Socher 在 RNN 模型的基础上提出了更加符合文本分类要求的长短时记忆模型（Long Short-Term Memory, LSTM）[27]；Brueckner 则认为应该同时考虑文本的上下文信息，因此在 LSTM 的基础上，提出双向 LSTM，进一步提高了文本分类的准确率[28]。何炎祥等人构建了情感分类模型 EMCNN，在模型的特征中首次使用了微博的表情符，取得了很好的分类效果[29]；梁军等人为了实现文本特征的自动选择，使用了递归自编码模型（Recursive Autoencoders, RAE），结果表明使用这种模型提取的特征可以提高情感分类准确率[30]。

1.3 本文研究内容概述

本文拟获取公开的政协提案数据，采用机器学习、深度学习、自然语言处理等相关技术，对政协提案及相关舆情进行挖掘分析，为政协委员提供信息技术支持。

具体工作包括以下几个方面：

(1) 对政协提案划分主题并提取关键词。编写政协网站的爬虫程序，获取公开的政协提案数据；对政协提案数据划分主题，将主题相同的提案划分为同一类;设计关键词提取算法，从每类主题中提取出一组能够精确表示主题内容的关键词。

(2) 情感分类模型的设计、训练、数据标签预测。分析微博网页结构，开发微博爬虫程序，采集与每个关键词相关的微博数据并保存为结构化格式；采用多种算法设计不同的情感分类模型，使用部分打标数据训练模型并在测试数据集上进行测试，最终选择测试准确率高的模型预测所有未标注数据的标签。

(3) 从不同角度对政协提案的相关舆情进行统计并可视化。从关注度演进趋势方面分析民众对每个主题的关注度变化；从关注度大小方面分析政协委员和民众对每个主题的总体关注度；从情感演进趋势方面分析民众对每个主题的情感变化；从情感倾向方面分析民众对每个主题的总体情感倾向。

1.4 论文结构安排

本文总共包括六章，每一章的研究内容如下：

第 1 章为绪论部分。主要介绍了本文研究的背景和意义、国内外的相关研究现状、本文的主要研究内容和结构安排；

第 2 章为相关技术介绍部分。主要介绍了中文分词、词性标注、文本表示、特征选择、聚类、舆情分析、传统机器学习和深度学习等几个方面的相关技术；

第 3 章为需求分析和系统总体设计部分。首先展示了本文提出的政协提案相关舆情分析系统的设计思想和总体设计，之后简单介绍了系统各部分的作用和实现方法，并对简单模块的设计思想和实现方法做了具体论述；

第 4 章为相关模块的具体设计部分。主要对关键词提取模块、情感分类模块进行介绍，详细说明了每个模块在每个步骤中的实现方法和结果；

第 5 章为实验结果分析部分。主要对本文提出的两种关键词提取算法的效果进行对比，并从不同方面分析了政协提案的相关舆情；

第 6 章为结论部分。主要总结了本文的研究成果，并对后续工作的方向进行了展望。

2 相关技术介绍

为了在后续章节中更清楚的阐述本文的研究内容，本章对本文中使用到的相关技术进行介绍。第一节介绍文本特征工程方面的相关技术；第二节介绍常用的传统的聚类方法；第三节介绍舆情分析领域中的情感分析技术；第四节介绍情感分析领域中常用的传统的机器学习算法；第五节介绍深度学习算法；第六节为本章小结。

2.1 文本特征工程

本文的数据类型均为文本数据，涉及到很多文本数据的处理工作，因此使用文本特征工程方面的相关技术对本文的数据进行处理，文本特征工程的一般步骤包括文本预处理、特征选择和文本表示，下面介绍这三个方面的相关技术。

2.1.1 文本预处理

文本预处理的对象可以是任何语言的文本数据，具体方法是使用分词、词性标注和停用词过滤等技术将原始的文本数据进行处理，使处理后的文本数据为结构化格式，从而可以用于后续研究。

（1）分词

由于英文等其他多种语言的文本数据的词与词之间由空格分隔，因此可以根据空格对英文文本数据等进行分词，而中文文本的词与词之间没有分隔符，因此中文分词首先需要识别每个词的边界，然后将由字组成的句子以词为单位进行切割，作为后续其他自然语言处理任务的基础。中文分词在很多领域都具有广泛的应用，包括自动摘要、文本分类和机器翻译等[31]。目前主要有两种常用的分词方法，分别是基于词典的算法和基于统计的机器学习算法[32]。

基于词典的分词算法通常是遍历文本数据中的所有不同长度的字符串，将其与词典中的词进行匹配，如果匹配成功，则说明目前的字符串是一个词。基于词典的分词算法分词速度快、简单高效，在很长一段时间内研究者都在对基于词典的方法进行各种优化，因此具有广泛的应用。但是由于不可能创建一个包含所有中文词语的词典，因此这种方法对于词典中未包含的词(未登录词)分词效果不好。

基于统计的机器学习算法包括有监督学习算法和无监督学习算法两种类型。有监督学习方法是使用机器学习算法对人工分词的大量文本数据进行训练，将训

练好的模型用于未知的文本数据，实现对未知文本数据的分词；无监督学习方法根据字符串在大量的文本语料中出现的频率来确定该字符串是否可以构成词。基于统计的机器学习算法可以解决词典法不能识别未登录词的缺点，可以取得比词典法更好的效果，但是基于统计的机器学习算法也有很多缺点，其中有监督学习方法需要大量的人工分词数据训练分词模型，造成大量的人力消耗，无监督学习方法在设置共现字出现频率的阈值上存在很大的主观性。

现有研究成果中有很多成熟的分词工具，例如常用的分词工具有 BosonNLP[33]、哈工大语言云[34]、结巴分词[35]等。

(2) 词性标注

词性标注是指使用某种规则或算法标注出文本数据中每个词的词性。目前常用的词性标注算法主要分为以下两种：基于规则的词性标注算法和基于统计机器学习的词性标注算法。

基于规则的词性标注算法是一种匹配算法，将文本数据中的每个词与词性词典中的词进行匹配，从而得到每个词的所有词性，之后再结合词语的上下文语境和语法规则消除歧义词性，最终标注出每个词的唯一词性[36]。这种方法简单易懂，但是需要构造完备的词性词典，消耗大量的人力。

基于统计机器学习的词性标注算法计算句子中每个词的所有可能的词性，从而使每个句子对应多个词性序列，然后使用特定的算法计算每个词性序列的分数，将所有的分数从大到小排序，最终选择分数最高的词性序列作为每个句子的词性序列。词性标注任务中常用的机器学习算法有隐马尔科夫模型(Hidden Markov Model, HMM)[37]、条件随机场(Conditional Random Field, CRF)[38]、最大熵马尔可夫模型(Maximum Entropy Markov Model, MEMM)[39]、循环神经网络(Recurrent Neural Network, RNN)[26]等。

(3) 停用词过滤

停用词一般指在文本中出现非常频繁的词或没有实际意义的词，如：“我”、“就”、“的”等词。这两种类型的词在自然语言处理任务中不仅没有任何意义，而且还会造成干扰。目前常用的去除停用词的方法有两种，分别是基于停用词表的方法和基于词性标注的方法[40]。

基于停用词表的方法是一种匹配式方法，主要是循环遍历分词后的文本数据中的每个词，将文本中的词与停用词表中的词进行匹配，如果匹配成功，则说明这个词是停用词，将其从文本数据中删除，否则保留。停用词表可以根据自己的文本数据构建，也可以使用公开的停用词表，一般情况是以公开的停用词表为基础，再添加自己的文本数据中的停用词。

基于词性标注的方法首先使用词性标注算法来标注待分析的文本数据中每个

词的词性，然后根据任务需求去除某些词性的词，通常情况下，去除其中的助词、量词、连词、介词、语气词等。

2.1.2 特征选择方法

在机器学习应用中，数据通常包含多个特征属性，其中可能存在冗余特征或不重要的特征，这会造成如下问题：特征数量越多，越容易造成维数灾难，模型训练时间就越长；特征数量越多，越可能存在冗余特征，对模型训练造成干扰，降低模型的泛化能力。因此需要对数据进行特征选择。

特征选择是指针对待研究的数据的特点和特定的任务需求，选择合适的特征选择算法计算数据中每个特征的权重，并设置合适的权重阈值，保留权重大于阈值的特征，剔除权重小于阈值的特征。经过特征选择，可以剔除数据中不相关的特征或冗余特征，从而减少特征数量，提高模型的训练速度、准确率以及泛化能力。常用的特征选择方法主要有词频-逆文档频率(Term Frequency–inverse Document Frequency, TF-IDF)、卡方检验等，下面依次介绍这两种特征选择方法。

(1) TF-IDF 算法

TF-IDF 算法的主要思想是如果某个特征在数据集中的一条数据中出现次数很多，在该数据集的其余数据中很少出现，则认为这个特征具有很好的区分能力，并且这个特征在该数据中的 TF-IDF 值大于在其余数据中的 TF-IDF 值。

TF-IDF 算法中的 TF 指词频，表示每个特征在每个样本数据中的出现次数；DF 指文档频率，表示某个特征在数据集中的所有数据中出现的频率，IDF 表示将 DF 的值取倒数，再将结果取对数。

假设语料库 C 中共有 D 条样本数据，则 TF-IDF 的公式如下：

$$tfidf_{i,j} = tf_{i,j} * idf_i \tag{2-1}$$

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2-2}$$

$$idf_i = log\frac{|D|}{|\{j: t_i \in d_j\}|} \tag{2-3}$$

其中$n_{i,j}$表示词t_i在样本数据d_j中的出现次数，$\sum_k n_{k,j}$表示样本数据d_j中所有词的出现次数之和，$|\{j: t_i \in d_j\}|$表示包含词t_i的样本数。

(2) 卡方检验

卡方检验首先提出原假设，通过计算卡方统计量并查找卡方检验表，以一定的概率相信原假设或相信备择假设。假设有两个分类变量 X 和 Y，每个变量分别有两

个特征，分别是$\{x_1,x_2\}$和$\{y_1,y_2\}$，其中 a、b、c、d 分别表示(x_1,y_1)、(x_1,y_2)、(x_2,y_1)、(x_2,y_2)的实际取值，A、B、C、D 分别表示在“X 与 Y 无关”的假设下计算的(x_1,y_1)、(x_1,y_2)、(x_2,y_1)、(x_2,y_2)的理论值。

在卡方检验中，首先提出原假设：“X 与 Y 无关”，计算卡方统计量的值并查找卡方检验表，从而精确的给出这个假设的可靠程度，卡方统计量的计算公式为：

$$\sum_{i=1}^{n}\frac{(x_i-E)^2}{E} \tag{2-4}$$

其中x_i表示理论值，E 表示实际值。

2.1.3 文本表示方法

在自然语言处理任务中，通常需要将文本数据交给计算机处理，由于计算机无法识别人类语言，因此我们需要使用文本表示方法将文本数据转化为向量形式，再输入计算机中进行处理。目前常用的文本表示方法包括以下两种：布尔编码表示和分布式表示。

布尔编码表示又称为 One-hot representation，就是将一个词用一个很长的向量来表示，向量的长度是将语料库中的所有样本中的词去重构成的词典的大小，向量的分量中只有一个位置的值为 1，其余位置的值均为 0，1 的位置表示该词在词典中的索引。布尔编码表示简单直观，但也存在一些缺点，第一，这种方法认为句子中的每个词都是孤立的，因此没有挖掘出词与词之间的语义信息；第二，当语料库中的样本数量较多导致词典较大时，向量会变得高维稀疏，容易造成维数灾难，不仅会影响聚类或分类效果，而且会造成很大的计算量并占用大量内存资源。

分布式表示又称为 Distributed representation，其基本思想是：通过对语料库中的大量文本数据进行训练，将其中的每个词用一个固定长度的短向量表示，向量的长度可以根据样本数量和任务需求的不同进行调整。目前使用最广泛的一种分布式表示方法是 Word2Vec[41]。

Word2Vec 是由 Google 在 2013 年提出的，其基本思想是以语料库中的文本数据作为输入，通过训练将每个词转换为固定长度的词向量，词向量中的数值不再是稀疏的 0，1 表示，而是浮点数。Word2Vec 共有两种训练算法模型，分别为 CBOW(Continuous Bag-of-Words Model)和 Skip-gram(Continuous Skip-gram Model)，下面依次介绍这两种训练算法模型。

CBOW 模型的结构如图 2-1 所示：

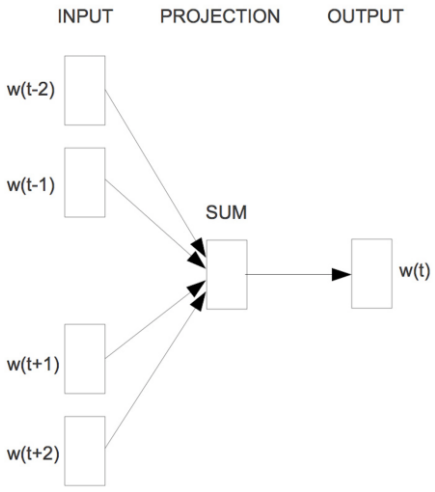


图 2-1 CBOW 模型结构图[41]

Figure 2-1 The structure of CBOW model[41]

由图 2-1 可知，CBOW 模型由三层不同的结构组成，分别是输入层、投影层和输出层。对于语料库中的每条样本数据，该模型是通过输入每个样本中某个中心词的前后几个词来预测该词出现的概率。输入层输入的是待训练词即中心词 w 的前后各 c 个词的词向量，其中 c 可以自己设定，以“今天的天气很晴朗”为例，将该句子分词后的结果为：“今天”、“的”、“天气”、“很”、“晴朗”，假设我们要得到“天气”的词向量，以 c 的值取 1 为例，“天气”的前后各 1 个词分别是“的”和“很”，则输入层输入这两个词的随机初始化的词向量；投影层将输入的词向量相加并输入到输出层；输出层则根据不同的结构采用不同的算法。

Skip-gram 模型与 CBOW 模型的原理相同，但网络结构正好相反。Skip-gram 模型的结构如图 2-2 所示：

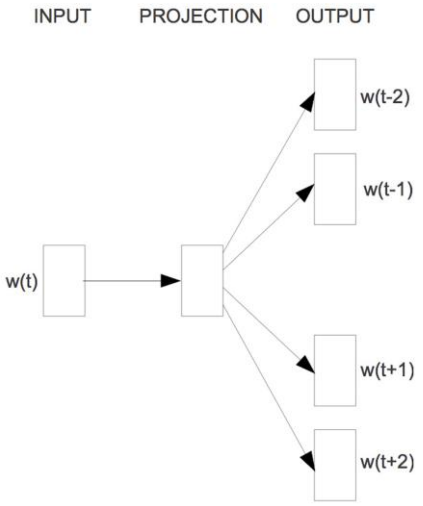


图 2-2 Skip-gram 模型结构图[41]

Figure 2-2 The structure of Skip-gram model[41]

由图 2-2 可知，与 CBOW 模型相同，Skip-gram 模型也由输入层、投影层和输出层构成。输入层是中心词 w 随机初始化后的词向量；为了与 CBOW 模型对比，投影层是一个恒等投影；输出层也是根据不同的结构采用不同的算法。

2.2 聚类方法

聚类是根据数据集中样本数据的特点将特征相同的样本聚为同一类，每一类称为一个簇。目前关于聚类技术的研究比较成熟，有很多不同的聚类算法。本节只具体介绍在本文中使用到的 K-means 聚类算法[42]。

K-means 聚类中的 k 表示将待分析的样本数据划分为 k 个簇，使每个簇中的样本均属于同一个类别，means 表示取每个簇中样本数据的均值作为该簇的质心，并用每个簇的质心作为对该簇的描述。

K-means 聚类算法的核心是确定 k 值的大小，通常需要选择合适的聚类效果评价指标，通过调节 k 值的大小，观察聚类评价指标的变化，选择使聚类评价指标最好的 k 值做为最终的 k 值。常用的评价指标有平均半径、平均直径、紧凑度、分离度、戴维森堡丁指数、轮廓系数等。

K-means 聚类算法的流程如下：

输入：样本集合 S，聚类的类别数 k，迭代轮数 N

输出：划分好的 k 个簇

1) 选择初始质心。从样本集合 S 中选择相对距离最远的 k 个样本点作为初始质心；

2) 划分样本。对于除质心以外的每个样本，计算样本与每个质心之间的距离，将样本划分到距离最近的质心所代表的簇中；

3) 更新质心。对每个簇中的样本取均值，作为新的 k 个质心，计算原来的 k 个质心和更新后的质心之间的距离，判断质心是否发生偏移；

4) 重复 2、3 步骤，直到达到迭代次数或质心不再偏移为止。

2.3 情感分析方法

情感分析通常是对语料库中的样本数据做情感分类，目的是自动判断样本数据的情感极性或强度。目前常用的情感分析方法包括基于情感词典的情感分类、基于传统的机器学习的情感分类和基于深度学习的情感分类。

2.3.1 基于情感词典的情感分类

基于情感词典的情感分类方法是一种匹配式方法，对于语料库中每一条分词后的样本数据，将样本中的每个词与词典中的词进行匹配，记录每个样本数据在情感词典中匹配成功的词，将这些词对应的强度值相加作为每个样本的情感强度。因此这种方法的核心是构建情感词典，一般是在通用情感词典的基础上添加领域词，中英文领域中常用的通用情感词典如表 2-1 所示：

表 2-1 常用的情感词典及简介
Table 2-1 Commonly used emotional dictionary and introduction

词典名称	语言种类	简介
WordNet	英语	同义词情感语义词典
SentiWordNet	英语	同义词情感词典，包括积极、消极、中立
General Inquirer	英语	国外最早的一个情感词典
HowNet	中英文	知网开发的情感词典
NTUSD	中英文繁体	一款包含中英文的情感词典
DUTIR	中文	大连理工大学开发的情感词典
BosonNLP	中文	Boson 中文情感词典

扩展通用情感词典的方法包括两种，分别是专家标注法和自动扩建法。专家标注法一般是语言学家根据专业知识对每个情感词进行人工标注，包括情感极性和强度的标注，但是这种方法构建的情感词典主观性较强，且构造过程要花费大量的时间和精力；自动扩建法一般是选择一个通用情感词典，将待分析的语料库中的词与词典中的词进行相似性计算，如果语料库中的词与词典中的某个词相似度大于给定的阈值，则认为该词是情感词，并将词典中的相似词的极性和强度赋给该词。

2.3.2 基于传统机器学习的情感分类

基于传统机器学习的情感分类方法需要使用少量的已标注数据训练模型并在测试集上测试，不断调整模型参数，使模型在测试集上的准确率尽可能提高。然后使用训练好的模型预测所有未标注数据的标签。

情感分类模型分为情感二分类和情感多分类。情感二分类一般根据样本数据的情感极性进行二分类标注，将积极情感的样本数据标注为 1，消极情感的样本数

据标注为 0。情感多分类根据样本数据的情感强度值的大小进行多分类标注，通常分为情感五分类或七分类等，以情感五分类为例，一般根据样本数据情感强度值的大小将其标注为 5，3，1，-1，-3，-5 等。

将打标后的数据使用机器学习分类算法进行训练，从而使用训练好的分类器预测所有未标注数据的标签。由于机器学习分类算法均为线性模型，不能很好的学习数据中的非线性特征，在复杂数据上表现不好，而且不能自动选择特征。为了得到更好的分类效果，通常对样本数据进行人工特征选择，保留重要的特征，剔除不重要的特征或干扰特征。

2.3.3 基于深度学习的情感分类

深度学习算法通常具有多层非线性网络结构，不仅可以学习数据中的非线性特征，在复杂数据上取得较好的效果，而且由于深度学习模型中复杂的非线性网络结构的存在，模型可以自动选择特征，从而减少了人为选择特征的主观性。

常用于搭建深度学习情感分类模型的算法有卷积神经网络(Convolutional Neural Networks, CNN)[43]、循环神经网络(Recurrent Neural Network, RNN)[25]、长短时记忆网络(Long Short Term Memory Network, LSTM)[26]等。卷积神经网络最初主要用于计算机视觉领域，通过不断提取图像的边缘信息来分析不同图像的特征，从而识别不同的图像或对图像分类，现在也应用于自然语言处理领域；由于做情感分类的样本数据均为文本数据，数据的上下文之间存在语义关系，而循环神经网络善于处理时序数据，因此常用循环神经网络做情感分类；长短时记忆网络是对循环神经网络的改进，本质上也是一种循环神经网络。

基于深度学习的情感分类模型的训练与基于传统机器学习的情感分类模型的训练过程类似。将标注数据输入深度学习模型中训练，调整模型参数，不断提高测试准确率，使模型达到最优，然后使用训练好的模型预测所有未标注数据的标签。

2.4 传统机器学习算法

下面介绍几种常用于情感分类任务的传统机器学习算法，包括朴素贝叶斯[44]47-48、逻辑回归[44]77-80 等。

2.4.1 朴素贝叶斯

朴素贝叶斯是一种基于样本数据的特征条件独立性假设与贝叶斯理论的分类

方法[44]47-48。已知样本数据集 T={$(x_1,y_1),(x_2,y_2),\cdots\cdots(x_n,y_n)$}，朴素贝叶斯需要根据已知的样本数据集学习联合概率分布 P(X,Y)，由于 P(X,Y)=P(Y)*P(X|Y)，因此学习联合概率分布 P(X,Y)可以分解为学习先验概率分布 P(Y)和条件概率分布 P(X|Y)。

先验概率分布可以根据样本数据统计得到，公式为：

$$P(Y=c_k)，\ k=1,2,\cdots,K \tag{2-5}$$

条件概率分布的公式为：

$$P(X=x|Y=c_k)=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k\right)，\ k=1,2,\cdots,K \tag{2-6}$$

由于每个样本数据具有 n 维特征，每个特征互相不独立，因此条件概率分布具有指数级数量的参数，直接估计是不可行的。所以假设条件概率分布中的特征相互独立，公式为：

$$P(X=x|Y=c_k)=\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}|Y=c_k\right)，\ k=1,2,\cdots,K \tag{2-7}$$

朴素贝叶斯算法在分类时，使用训练好的模型计算输入数据 x 的后验概率分布$P(Y=c_k|X=x)$，后验概率最大的类就是输入数据 x 的类别。因此朴素贝叶斯算法的分类模型的公式为：

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)P(X=x|Y=c_k)}{\sum_k P(Y=c_k)\,P(X=x|Y=c_k)} \tag{2-8}$$

经化简后，可以简化为：

$$y=arg\max_{c_k}P(Y=c_k)\prod_j P\left(X^{(j)}=x^{(j)}\middle|Y=c_k\right) \tag{2-9}$$

2.4.2 逻辑回归

逻辑回归是一种简单高效的分类算法，在很多领域具有重要的应用，如判断某封邮件是否是垃圾邮件，判断某个用户是否是潜在用户等。逻辑回归可分为二项逻辑回归模型和多项逻辑回归模型，可以分别实现数据的二分类和多分类[44]77-80。

二项逻辑回归模型用于对样本数据做二分类，条件概率分布的公式如下：

$$P(Y=1|x)=\frac{exp(w*x+b)}{1+exp(w*x+b)} \tag{2-10}$$

$$P(Y=0|x)=\frac{1}{1+exp(w*x+b)} \tag{2-11}$$

其中 x 为输入的样本数据，Y 是样本数据的类别。

二项逻辑回归模型通过分别计算输入样本数据 x 属于 0 和 1 的概率，将样本数据 x 的类别判定为概率值大的那一类。

多项逻辑回归模型由二项逻辑回归模型扩展得到，用于对样本数据做多分类，假设某个语料库中所有的样本数据包含 K 个类别，条件概率分布的公式如下：

$$P(Y=k|x)=\frac{exp(w_k*x)}{1+\sum_1^{K-1}exp(w_k*x)},\ k=1,2,\dots,K-1 \tag{2-12}$$

$$P(Y=K|x)=\frac{1}{1+\sum_1^{K-1}exp(w_k*x)} \tag{2-13}$$

2.5 深度学习算法

下面介绍几种常用于情感分类的深度学习算法，由于情感分类的数据均为文本数据，因此主要介绍常用于文本数据的深度学习算法模型，包括循环神经网络(Recurrent Neural Network, RNN)[26]、长短时记忆网络(Long Short Term Memory Network, LSTM)[27]、双向长短时记忆网络(Bi-directional LSTM,Bi-LSTM)[28]，其中 LSTM 是对 RNN 的改进，Bi-LSTM 是对 LSTM 的改进。

2.5.1 循环神经网络

循环神经网络是一种用于处理时间序列数据的神经网络模型[26]，常见的时间序列数据如语音数据、文本数据等均适用于使用循环神经网络进行处理。

循环神经网络模型如图 2-3 所示：

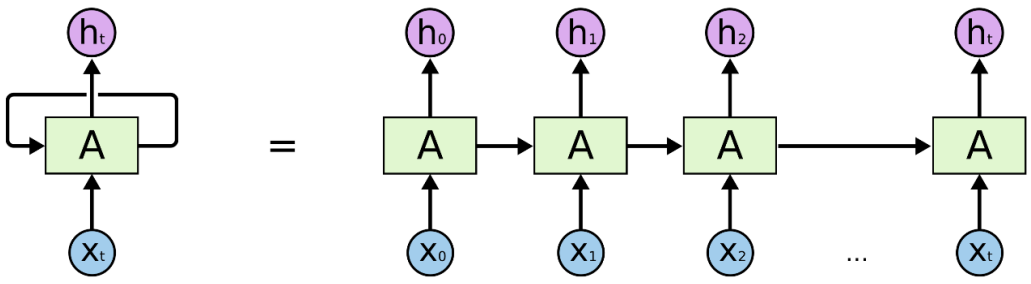


图 2-3 循环神经网络结构图

Figure 2-3 Recurrent neural network structure

由图 2-3 可知，等号左边表示循环神经网络的递归结构，等号右边表示循环神经网络按时间展开后的结果。图中x_i表示每个时刻的输入数据，h_i表示每个时刻的输出特征，其中h_t表示整个序列的最终特征，一般将该特征输入到后续网络结构中进行计算。

虽然循环神经网络可以处理时序数据，但在循环神经网络的训练过程中存在一系列问题，如在循环神经网络的前向传播过程中存在长期依赖问题；在反向传播过程中容易造成梯度消失和梯度爆炸问题，这些问题会影响神经网络的性能。

2.5.2 长短时记忆网络

为了解决循环神经网络面临的长期依赖、梯度爆炸和梯度消失问题，人们研究出多种循环神经网络的变体，其中应用最为广泛的是长短时记忆网络[27]，它通过引入三个门控制单元，希望在不同时刻有选择的输入部分有用信息或有选择的丢弃部分无用信息，从而解决循环神经网络存在的问题。

长短时记忆网络中增加的三个门结构单元分别是输入门、遗忘门和输出门。输入门并不是将当前的全部信息加入内部状态中，而是有选择的加入部分有用信息；遗忘门控制上一时刻的内部状态中有多少无用信息需要丢弃；输出门控制当前的内部状态中有多少信息需要输出，作为本时刻的输出状态。

长短时记忆网络可以有效地建立长距离的时序依赖关系，目前已经在时序预测、自然语言处理等领域取得了很好的效果。

2.5.3 双向长短时记忆网络

长短时记忆网络虽然解决了循环神经网络存在的长期依赖、梯度爆炸和梯度消失问题，但是处理时序数据的顺序是从左到右按照数据的正常阅读顺序输入模型中的，没有实现将数据从右往左输入模型中提取更加全面的信息。而双向长短时记忆网络实现了将数据从右往左输入模型中来提取更全面的信息[28]。

双向长短时记忆网络的网络结构如图 2-4 所示，图中的x_i表示每个时刻的输入数据，y_i表示每个时刻的输出特征，S_0表示从时序数据的起始位置开始将数据输入到模型中，S_0'表示从时序数据的终止位置开始将数据输入到模型中，将同一时刻从起始位置开始输入的数据产生的特征与从终止位置开始输入的数据产生的特征相加，作为该时刻的输出特征。

双向长短时记忆网络同时考虑了时序数据的正向语义和反向语义，对数据的特征提取更加全面精确，因此可以提高模型的准确率。

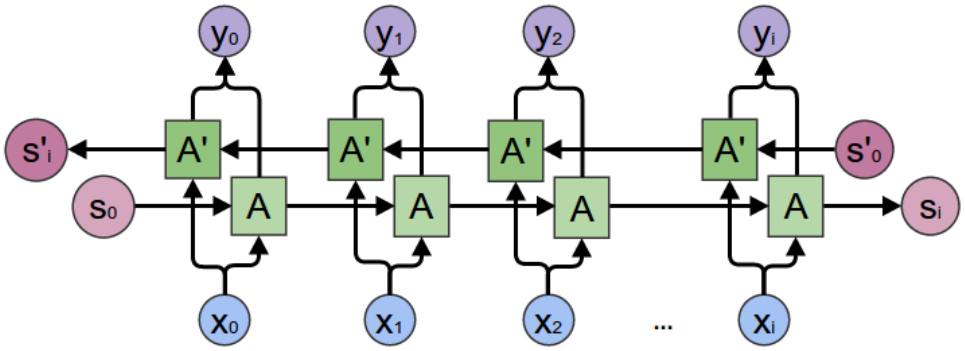


图 2-4 双向 LSTM 结构图

Figure 2-4 Bidirectional LSTM structure

2.6 本章小结

本章对本文研究所用到的相关技术进行了详细介绍。第一节介绍了文本预处理技术、常用的特征选择方法、文本表示方法等；第二节主要介绍了本文研究使用的 K-means 聚类算法的原理；第三节介绍了三种现有的情感分析技术；第四节介绍了几种常用于情感分析领域的传统机器学习算法；第五节介绍了几种常用于做情感分析的时序模型。本章介绍的相关技术，既是自然语言处理领域的一些关键技术，也是本文系统设计和实现的基础。

3 需求分析和系统总体设计

本章主要介绍本文研究的政协提案舆情分析系统的整体设计和系统中每个模块的整体结构。本章分为 8 部分，第一部分介绍本文所研究的问题，第二部分介绍政协提案舆情分析系统的整体设计，第三部分介绍政协提案爬虫程序的设计和提案数据的获取，第四部分介绍对政协提案划分主题的方法和关键词提取方法，第五部分介绍微博爬虫程序的设计难点、相应的解决方案和爬虫模型的结构，第六部分介绍情感分类模型的设计方法，第七部分介绍舆情分析方法，第八部分为本章小结。

3.1 需求分析

3.1.1 背景介绍

全国政协提案是我国政治制度非常重要的机制之一，在互联网络时代，利用技术手段挖掘提案的热点主题和相关舆情具有现实意义。可以为政协委员从整体上把握提案的关注点提供技术信息的参考，从而节约政协委员以后形成新提案的时间；可以发现相应的社情民意，从而使得政协委员能够把握提案的舆情效果，节约政协委员形成新的提案的时间和精力。

目前，关于提案的热点主题发现和采用技术手段对热点主题进行舆情统计的研究尚未见到。对政协提案，目前包括以下几个方面的相关研究：研究政协提案在政府决策中的作用[2]；研究政协提案的采纳和办理程度[3]；研究政协提案的信息化管理系统的设计[4]等。这些研究均未对提案的热点主题和相关舆情进行挖掘分析。

本文拟获取公开的政协提案数据，对政协提案数据进行挖掘分析，为政协委员提供信息技术支持。具体如下：将相同主题的提案划分为同一类别，从而发现政协委员关注的热点主题；从每个主题中提取一组关键词，使用微博爬虫程序获取每个关键词的微博数据，通过分析微博数据得到人们对每个主题的关注度和情感变化。

3.1.2 设计内容

在本文实验中，需要设计一个提案的主题发现及相关舆情分析系统。具体如下：

(1) 数据来源。分析对象是政协提案，因此需要获取相应的政协提案数据。

(2) 提案主题划分与关键词提取。所有提案可能会属于某几个主题，因此需要

使用主题划分方法将相同主题的提案划分为同一类，从每类主题中提取关键词。

(3) 获取与关键词相关的微博数据。通过分析微博网页结构开发一套微博爬虫程序，爬取每个关键词对应的微博数据。

(4) 舆情分析。使用爬取的微博数据分析人们对每个主题的关注度大小和演进趋势，设计情感分类模型，分析人们对每个主题的情感倾向和演进趋势。

3.2 整体设计步骤

为了实现政协提案舆情分析系统的设计需求，下面给出系统的整体设计。整体设计如图 3-1 所示：

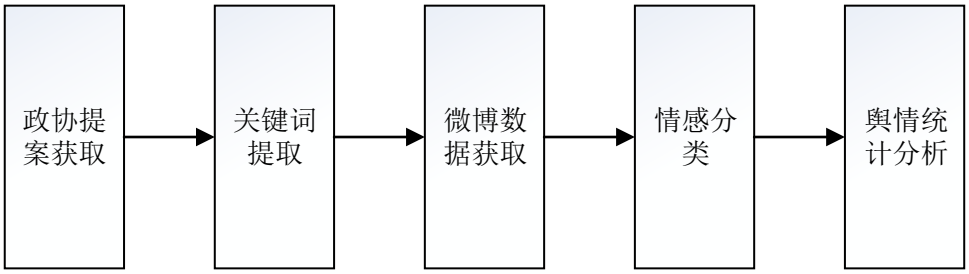


图 3-1 系统的整体设计

Figure 3-1 The overall design of the system

从图 3-1 可以看出政协提案及其相关舆情的分析系统由五个模块组成：

(1) 政协提案获取模块。编写网络爬虫程序，从北京市政协网站采集 2018 年的政协提案数据，作为待分析对象。

(2) 关键词提取模块。如果从每件提案中分别提取关键词，数量很多，会导致关键词冗余(相同或相似)，致使后续进行大量的冗余处理。因此先对提案划分主题，再从主题中提取关键词。

(3) 微博数据获取模块。分析微博网站的网页结构，开发微博爬虫程序用于采集与提取出的关键词相关的微博数据。

(4) 情感分类模块。设计基于双向 LSTM 的情感分类模型，使用部分打标数据训练模型并在测试集上进行测试，调参得到最优模型，使用最优模型预测所有未标注数据的标签。

(5) 舆情统计分析模块。从关注度分析和情感分析两方面展开，统计每个主题在不同时间段的微博数量和每个主题的微博数据总量，分析人们的关注度演进趋势和关注度大小；统计每个主题在不同时间段的不同情感极性的微博数量和每个主题的不同情感极性的微博数据总量，分析人们的情感演进趋势和总体情感倾向。

下面依次介绍每个模块的整体结构设计。

3.3 政协提案获取

全国各级政协委员每年都会提出数以百计的提案并公布在各级政协网站上，由于政协网站没有提供批量下载功能，因此本文设计爬虫程序实现批量获取提案。

本文通过分析北京市政协网站的网页结构，开发了北京市政协网站的爬虫程序并实现了北京市政协提案数据的自动采集。这种方法很容易扩展到其他各级政协网站并获取相应的政协提案。

爬虫模块的流程图如图 3-2 所示：

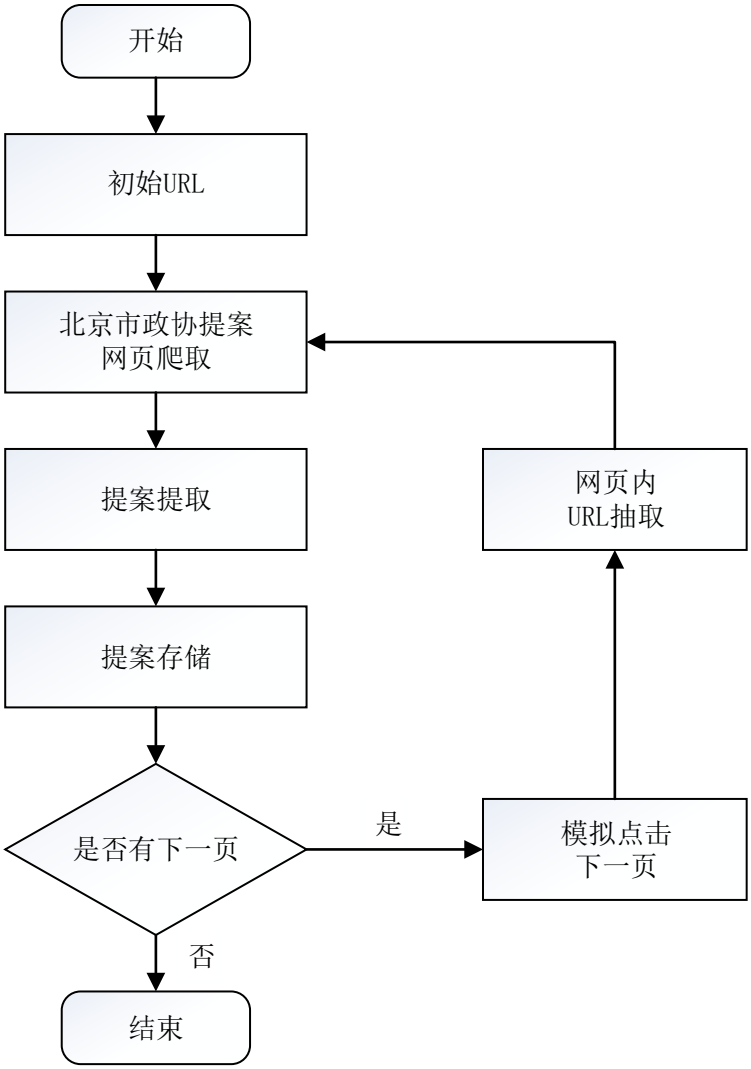


图 3-2 政协提案爬虫模块设计图

Figure 3-2 Design of CPPCC proposal reptile module

由图 3-2 可知，首先人工获取政协网站中第一页政协提案的 URL，输入到爬虫程序中，之后实现全自动的政协提案数据采集，直到最后一页政协提案采集完成后程序终止。

本文使用政协网站爬虫程序爬取了 2018 年北京市的全部政协提案数据，总共 798 件，并将每一件提案单独按 txt 格式保存，部分政协提案的保存结果如图 3-3 所示：

图 3-3 政协提案保存格式

Figure 3-3 CPPCC proposal save format

3.4 关键词提取

使用政协网站爬虫程序获取到北京市的政协提案数据后，需要从提案中提取关键词。由于政协提案数量众多，如果从每件提案中分别提取关键词，则提取出的关键词数量也很多，会导致关键词冗余，致使后续需要人工进行冗余处理。因此本文先对提案划分主题，将主题相同的提案划分为同一类，再从每类主题中提取关键词。

对获取的公开政协提案划分主题并从中提取关键词的流程如图 3-4 所示，可以看出，对政协提案划分主题并提取关键词的步骤包括数据预处理、主题划分和关键词提取。本节简要介绍三个步骤的作用和实现方法，具体的实现步骤和细节在第四章展开介绍。

(1) 数据预处理

数据预处理是文本数据处理的关键步骤，是后续研究的基础。本文对政协提案数据做以下预处理：分词、词性标注和停用词过滤。分词是指将由字组成的句子以

词为单位进行切割，本文使用改进后的结巴分词工具对政协提案数据进行分词处理；词性标注是指使用特定的规则或算法确定文本中每个词的词性并对其进行标注，本文使用结巴分词中的词性标注功能对政协提案中的每个词标注词性；停用词过滤是指去除所有文本中都频繁出现的词和没有实际意义的词，本文使用改进后的百度停用词表去除政协提案中的停用词。

(2) 主题划分

主题划分是指使用特定的算法将主题相同的政协提案数据划分为同一类，是后续关键词提取的基础。本文对政协提案的主题划分过程包括两个步骤：政协提案的向量化表示和聚类。在政协提案的向量化表示方面考虑两点：使用 TF-IDF 算法对政协提案中的词进行特征选择；充分考虑政协提案的结构，对标题和正文赋予不同的权重。综合考虑这两点，实现了政协提案的向量化表示。将向量化后的政协提案使用 K-means 聚类，将相同主题的提案聚为同一类。

(3) 关键词提取

关键词提取是指从文本或文本集中提取出一组可以精炼地表示文本或文本集的词语。本文从聚类形成的每个主题中提取关键词，因此属于从文本集合中提取关键词。本文设计了两种关键词提取算法，一种考虑每类主题中词的权重和词频，另一种考虑每类主题中词的权重、词频和表达能力。分别使用两种关键词提取算法从每类主题中提取出了两组关键词，并设计实验对比了两组关键词的有效性。

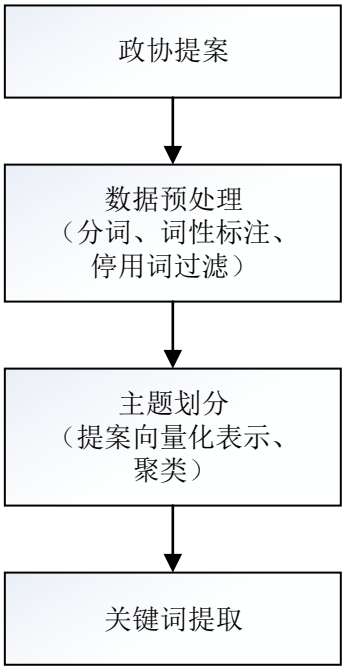


图 3-4 对政协提案划分主题并从主题中提取关键词

Figure 3-4 Divide the topic of the CPPCC proposal and extract keywords from the topic

3.5 微博数据获取

从每个主题中提取出能够精炼地表示主题内容的关键词后，需要获取与关键词相关的舆情数据。本文使用从每个主题中提取出的关键词采集微博舆情数据，由于微博网站没有提供批量获取微博数据的功能，因此通过分析微博网页结构开发微博爬虫程序，使用微博爬虫程序实现微博数据的批量采集。

3.5.1 难点及解决方法

微博是目前国内最大的社交媒体，每天都会有很多人从微博上采集数据，因此微博为了保证内部服务器的正常运行，对微博网页数据爬取设置了很多限制，目前已经形成了严格的反爬虫机制，本文在微博数据爬取的过程中遇到了以下几个问题：

(1) 用户访问限制。经过实验测试，在不登录微博的情况下，微博网站只能返回部分数据；

(2) IP 访问量限制。同一个 IP 在短时间内持续访问微博网站，微博网站可能会封禁该 IP，且短期内不能在该 IP 下登录微博；

(3) 用户访问量限制。在短时间内使用同一个账号频繁地登录微博，微博网站会增加输入验证码验证的步骤，而验证码的自动识别技术目前尚不成熟，每次手动输入验证码又不可行。

针对以上三个问题，分别采用如下的解决方法：

(1) 基于 Selenium 中的函数实现微博模拟登录，解决用户访问限制。

实现微博网站的模拟登陆是微博爬虫程序中的关键步骤，只有登录微博才能从微博网站中采集大量数据。本文使用 Selenium 库中的 Webdriver.Chrome()对象调用 find_element_by_css_selector()函数实现自动输入微博用户名和密码，并模拟点击微博的“登录”按钮，实现微博模拟登录。

(2) 使用多线程间接解决 IP 访问量限制。

通常情况下，解决 IP 访问量限制的方法是设置 IP 代理池，IP 代理池中有大量可用的 IP 地址，每次访问微博网页都使用不同的 IP 地址。然而爬取微博数据对代理 IP 的质量要求较高，开源的 IP 代理池中可用的 IP 很少，因此为了保证较快的微博数据采集速度，同时避免 IP 被封禁，本文使用 Scrapy 框架实现微博数据的多线程采集，但是设置相对较大的微博网页刷新时间间隔。

(3) 使用多个微博账号解决用户访问量限制。

通过注册多个账号，定时更换账号登录微博进行微博数据的爬取，可以有效的

减少输入验证码的次数。

3.5.2 爬虫整体设计

综合考虑微博数据采集中存在的三个难点，本文开发了微博爬虫程序，爬虫模块的结构如图 3-5 所示：

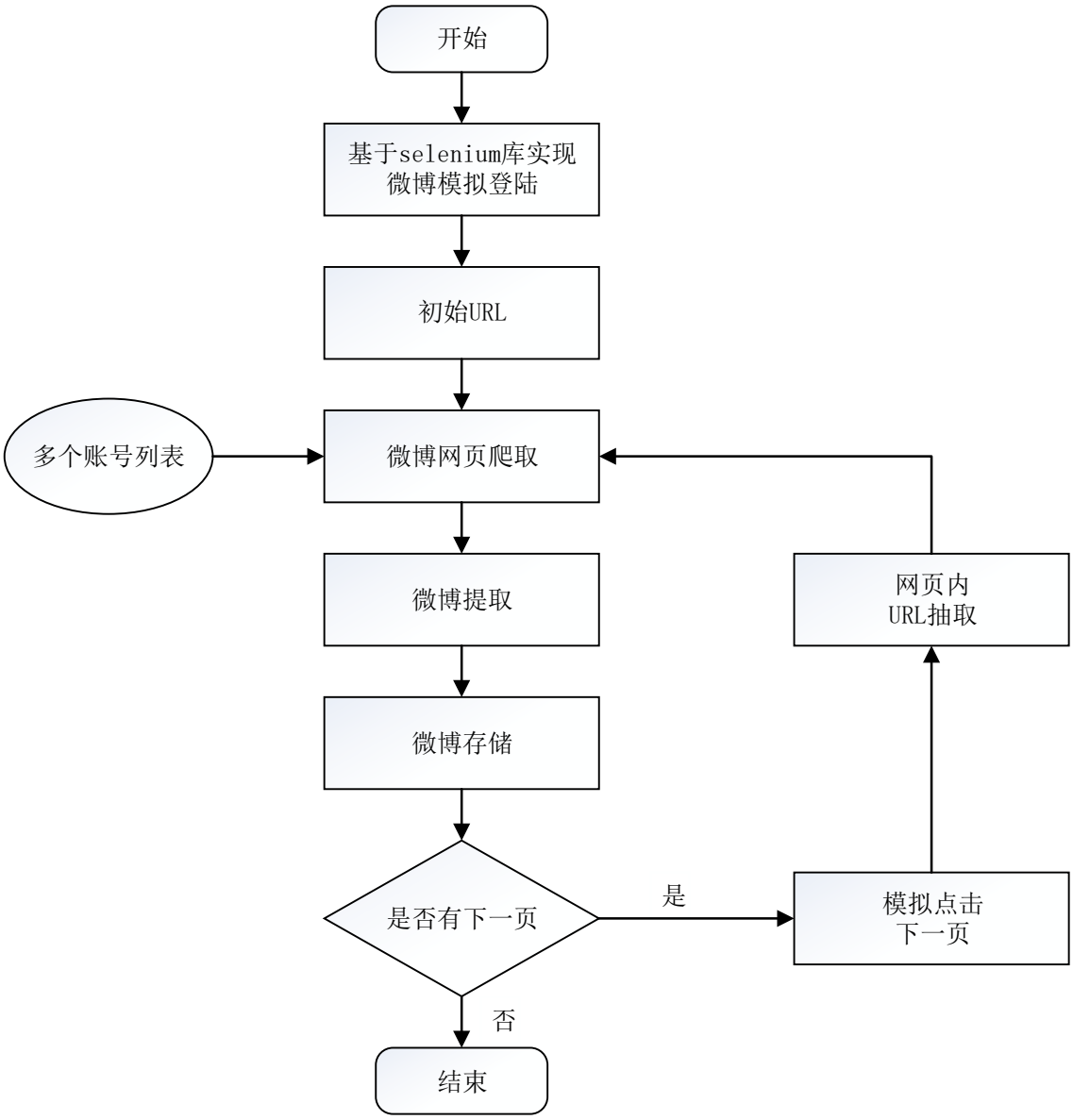


图 3-5 微博爬虫模块设计图

Figure 3-5 Design of Weibo reptile module

在图 3-5 中，第一步基于 Selenium 库中的函数实现模拟登陆微博；第二步通过爬虫程序中内嵌的初始 URL 获取微博网页，由于使用了 Scrapy 框架且有多个账

号轮流使用，因此减少了 IP 封禁和人工输入验证码的可能性；第三步分析微博网页结构并从中获取微博内容、发布微博的时间、微博评论等相关信息；第四步将获取的微博相关信息保存到本地的 Excel 表格中；第五步判断获取的微博网页中是否有“下一页”按钮，如果没有则说明已经爬取到了最后一页，程序自动终止，否则模拟点击“下一页”按钮，获取到下一页的 URL，分析下一页的新网页结构并采集相关微博数据。

使用本文开发的微博爬虫程序采集每个关键词在 2018 年的舆情数据，并将数据按关键词保存为结构化格式。

3.6 情感分类模型设计

使用微博爬虫程序获取微博舆情数据之后，需要标注每条微博数据的情感极性。由于人工对所有的微博数据标注极性会耗费大量时间和精力，在数据量特别大的情况下也是不现实的。因此本文人工标注少量数据，设计情感分类模型并使用标注数据训练模型，再使用训练好的模型预测所有未标注微博数据的情感极性。对微博数据进行情感分类的流程图如图 3-6 所示：

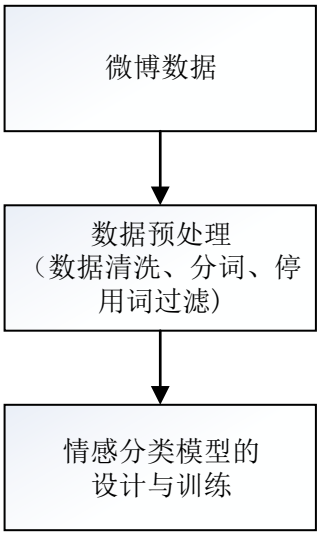


图 3-6 情感分类流程图

Figure 3-6 Emotional classification flow chart

由图 3-6 可知，对微博数据进行情感分类的步骤包括数据预处理和情感分类模型的设计与训练。本节只简单介绍这两个步骤的作用和实现方法，具体的实现步骤和细节在第四章中展开介绍。

(1) 数据预处理

数据预处理是指将原始的微博数据转化为情感分类模型可以使用的结构化格式，微博数据的预处理效果好坏在很大程度上会影响情感分类模型的准确率。本文对微博数据做以下预处理：数据清洗、分词和停用词过滤。数据清洗是指从原始微博数据中删除<p>、
等干扰字符和表情符、标点符单独构成的微博、采集失败的空微博等无效微博，从而筛选出纯文本微博数据用于情感分类模型的训练与预测；分词、停用词过滤的作用和实现方法与政协提案中的分词、停用词过滤的相关步骤相同，因此不再赘述。

(2) 情感分类模型的设计与训练

本文基于双向LSTM搭建情感分类模型，对部分微博数据人工标注情感极性，方法是：逐一阅读每条微博数据，根据微博数据的语义判断该句子表达的情感，正面情感的微博数据标 1，负面情感的微博数据标 0，将人工无法判断极性的微博数据舍弃。将人工标注的微博数据随机划分为训练集、验证集和测试集，其中训练集、验证集和测试集的比例为 8:1:1，然后将训练集输入到情感分类模型中训练，验证集输入情感分类模型中用于辅助训练，不断调节模型参数，在不发生过拟合的情况下，使模型在训练集和验证集上的准确率均达到相对最高，然后在测试集上进行测试。最后将训练好的情感分类模型用于所有未标注数据的情感极性预测。

3.7 舆情统计分析

使用情感分类模型预测出所有未标注数据的情感极性后，需要对数据进行统计，实现政协提案的相关舆情分析。政协提案的相关舆情主要从两个方面展开分析：关注度和情感。每个方面各从两个角度展开，分别是关注度大小和演进趋势分析、情感倾向和演进趋势分析。下面从理论上介绍本文的做法。

(1) 关注度大小分析

关注度大小分析指在政协提案提出后，对每个主题在 2018 年 2 月-2018 年 12 月的微博数量和每个主题中的政协提案数量进行对比，通过分析两者的对比结果得出民众和政协委员对每个主题的关注度异同。

(2) 关注度演进趋势分析

关注度演进趋势分析通过统计每个主题在不同时间段的微博数量，绘制每个主题的微博数量随时间的变化曲线，通过分析微博数量变化曲线得出民众对每个主题的关注度变化。

(3) 情感倾向分析

情感倾向分析指在政协提案提出后，对每个主题在 2018 年 2 月-2018 年 12 月的不同情感极性的微博数量进行统计，绘制每个主题不同情感极性的微博数量柱

状图和微博比例柱状图，从而得出民众对每个主题的总体情感倾向大小和比例。

(4) 情感演进趋势分析

情感演进趋势分析通过统计每个主题在不同时间段的不同情感极性的微博数量，绘制每个主题不同情感极性的微博数量随时间的变化曲线，通过分析不同情感极性的微博数量变化曲线得出民众对每个主题的不同情感变化。

3.8 本章小节

本章首先介绍了本文的研究问题，然后介绍了政协提案舆情分析系统的整体设计并简要介绍了每个模块的作用，之后详细介绍了每个模块的作用和实现方法，依次介绍了政协提案网站的爬虫设计，对爬取的政协提案划分主题并提取关键词的方法，微博爬虫程序的整体设计和在爬取过程中遇到的问题以及提出的解决方案，基于双向 LSTM 模型的情感分类模型的设计，多方面多角度的舆情分析方法。

4 相关模块的具体设计

由于关键词提取模块和情感分类模块涉及的算法较多，实现方法比较复杂，本文在第三章只介绍了这两个模块的作用和实现方法，本章详细介绍这两个模块具体的实现细节。本章分为六部分，第一部分介绍政协提案的数据预处理中遇到的问题及解决方法，第二部分介绍政协提案主题划分方法，第三部分介绍两种从每个主题的多件提案中提取关键词的方法，第四部分介绍微博数据预处理的具体实现方法和步骤，第五部分介绍情感分类模型的设计和模型每一层的功能，第六部分为本章小节。

4.1 关键词提取模块的数据预处理

本文在政协提案获取模块采用爬虫程序获取了政协提案数据，这些数据均为文本数据。本节对政协提案使用文本数据预处理的方法进行处理，预处理的步骤包括分词、词性标注和停用词过滤。

4.1.1 分词

分词就是识别每个词的边界，将由字组成的句子以词为单位进行切割。目前关于分词技术的研究比较成熟，有很多开源的分词工具可以使用。下面介绍本文对分词工具的选择和改进。

(1) 分词工具的选择和调用

通过对比几种常用的分词工具，最终选择结巴分词作为本文的分词工具。结巴分词具有以下优点：

1) 分词速度快，分词效果好，同一天之内可以无限次使用而没有次数限制。本文的政协提案数据和微博数据的数据量较大，尤其是微博数据更是达到了数十万条，因此选择一个速度快、效果好、没有使用次数限制的的分词工具可以提高实验效率和准确率；

2) 可以通过添加自定义词典进一步提高分词效果。由于政协提案数据的领域性较强，使用通用的分词工具对政协提案分词必然存在分错的情况，因此可以将分错的词语添加到自定义词典中提高分词准确率；

3) 具有词性标注功能。本文在提取关键词的过程中需要知道每个词的词性，

根据词性对不同的词语进行组合，使用结巴分词工具对政协提案分词后直接标注词性更方便，效率更高。

结巴分词工具提供了三种可选的分词模式，分别是全模式、精确模式和搜索引擎模式。基于本文的研究目标，选用精确模式进行分词，分词工具的接口函数如图 4-1 所示：

```
import jieba                                                    #第1行
jieba.load_userdict('/Users/lyj/Desktop/提案程序/sogou.txt')     #第2行
jieba.load_userdict('/Users/lyj/Desktop/提案程序/标题补充词典.txt') #第3行
words = jieba.cut(sentence,cut_all=False)                       #第4行
```

图 4-1 结巴分词的接口函数

Figure 4-1 The interface function of jieba

图 4-1 中的四行程序是结巴分词的主要接口函数程序。程序中的第一行表示将结巴分词库导入 Python 代码中；第二行和第三行是对结巴分词的改进，导入自定义词典，本文中自定义词典的构造方法和词典中的内容在后文中介绍；第四行是使用结巴分词工具对待分词的对象“sentence”进行分词，其中“cut_all=False”表示选择精确模式。

通用的结巴分词工具对政协提案分错的词较多。以提案数据“关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案”为例，分词结果如下表 4-1 所示：

表 4-1 结巴分词的分词结果

Table 4-1 Word segmentation result of jieba

提案数据	结巴分词的分词结果
关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案	'关于', '推进', '北京', '文化', '中心', '建设', '，', '打造', '首都', '文', '博文', '创', '设计', '孵化', '平台', '的', '提案'

从表 4-1 可以看出，使用通用的结巴分词工具对政协提案的分词结果中存在部分词语分错的情况，如“文化中心”，“文博文创”等词，因此需要对结巴分词添加自定义词典进行改进。

结巴分词的改进难度较大。结巴分词提供了添加自定义词典的功能，可以将分错的词语(又称“未登录词”)添加到自定义词典中，提高分词准确率。但实际操作中存在两个难点：

1) 虽然可以将政协提案中的所有未登录词添加到自定义词典中来提高分词准确率，但是由于政协提案的数据量较大，人工校验所有的分词结果会花费大量时间

和精力；

2) 当研究对象发生变化时，为了保证新的研究对象也具有较高的分词准确率，又需要校验新研究对象的分词结果。

因此每次更换研究对象都采用人工校验分词结果并将其中的未登录词添加到自定义词典中的方法在实际上是不可行的。

(2) 本文对结巴分词的改进方法

基于上述讨论，本文采用两种方式向结巴分词的自定义词典中添加未登录词以期尽可能提高分词准确率：

1) 基于对政协提案数据的分析，大多数政协提案数据的标题在很大程度上可以反映提案的主要内容，因此在时间和精力允许的范围内，本文将政协提案的标题中分错的词语添加到自定义词典中；

2) 搜狗输入法词库中收集了各个领域的词库，本文将搜狗输入法词库中可能与提案相关的词库添加到自定义词典中，添加了“十九大报告”、“戏曲曲艺”、“经济生活词汇”等 140 个词库。

使用添加自定义词典后的结巴分词工具对政协提案分词，将分词结果保存在本地文件中，仍以提案数据“关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案”为例，分词结果如下表 4-2 所示：

表 4-2 改进后的结巴分词的分词结果

Table 4-2 Word segmentation result of Improved jieba

提案数据	改进后结巴分词的分词结果
关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案	'关于', '推进', '北京', '文化中心', '建设', '，', '打造', '首都', '文博文创', '设计', '孵化', '平台', '的', '提案'

从表 4-2 中可以看出，通过添加自定义词典对结巴分词进行改进，提高了分词准确率。

4.1.2 词性标注

词性标注是指使用特定的规则或算法确定文本数据中每个词的词性并对其进行标注的过程，词性标注并不是数据预处理的必要步骤。由于本文在关键词提取的过程中需要对政协提案中的词按照词性进行组合，因此本文对政协提案数据进行词性标注。

现有的开源分词工具中有些提供了词性标注功能，本文选用结巴分词工具对政协提案数据进行词性标注，将标注词性的政协提案数据保存在本地文件中。以提

案数据“关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案”为例，词性标注结果如下表 4-3 所示：

表 4-3 结巴分词的词性标注结果
Table 4-3 Part-of-speech tagging result of jieba

提案数据	结巴分词的词性标注结果
关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案	('关于', 'p'), ('推进', 'v'), ('北京', 'ns'), ('文化中心', 'x'), ('建设', 'vn'), ('，', 'x'), ('打造', 'v'), ('首都', 'd'), ('文博文创', 'x'), ('设计', 'vn'), ('孵化', 'v'), ('平台', 'n'), ('的', 'uj'), ('提案', 'v')

4.1.3 停用词过滤

停用词一般指在所有文本中都频繁出现或对文本的语义表达没有实际影响的词。停用词过滤方法有两种：基于停用词表的方法和基于词性标注的方法，本文使用基于停用词表的方法。下面介绍停用词表的选择和改进。

(1) 停用词表的选择

目前公开的通用停用词表有很多，使用最广泛的有百度停用词表、哈工大停用词表、四川大学停用词表等，其中百度停用词表在新闻报道类的文本数据上表现最好[45]。由于政协提案数据与新闻报道类数据在语言表达上最相近，因此本文选用百度停用词表。

(2) 停用词表的改进

通用停用词表中只包含一些常见的停用词，而对于不同的研究任务，通常需要有针对性地对停用词表进行人工改进。本文在通用的百度停用词表的基础上添加了政协提案中的停用词，同时从互联网上下载了所有的表情符添加到百度停用词表中。

循环遍历政协提案中的每个词，将其与本文改进的停用词表中的词进行匹配，如果匹配成功，则说明是停用词，应从分词结果中删除，否则保留，将所有分词并去除停用词后的政协提案数据保存在本地文件中。为了说明去除停用词后的效果，以提案数据“关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案”为例，将该数据中的停用词过滤后的结果如下表 4-4 所示：

表 4-4 结巴分词的词性标注结果
Table 4-4 Results after removing the stop word

提案数据	结巴分词并过滤停用词
关于推进北京文化中心建设，打造首都文博文创设计孵化平台的提案	'推进', '北京', '文化中心', '建设', '打造', '首都', '文博文创', '设计', '孵化', '平台'

4.2 关键词提取模块的主题划分方法

主题划分是指将相同主题的政协提案划分为同一类。本文对政协提案划分主题的过程包括两个步骤：政协提案的向量化表示和 K-means 聚类。

4.2.1 提案向量化表示

(1) 提案向量化的考虑和解决方法

政协提案属于长文本数据，常用于长文本数据的向量化算法有隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)[46]、概率潜在语义分析(Probabilistic Latent Semantic Analysis, pLSA)[47]、潜在语义分析(Latent Semantic Analysis, LSA)[48]、Doc2vec[49]等。

这些方法直接将整个长文本训练为一个向量，虽然考虑了词与词、段落与段落之间的语义信息，但没有考虑文本的结构信息。本文要分析的政协提案是结构化文本，包括提案标题和提案正文，其中标题是对正文的精确提炼，因此标题和正文的重要性不同。

本文考虑政协提案的结构信息，采用如下方法对提案进行向量化表示：本文以词为粒度，对预处理后的政协提案中的词训练词向量；计算每个词的 TF-IDF 值，通过调参选择合适的 TF-IDF 阈值，剔除提案中小于 TF-IDF 阈值的词；分析政协提案的结构，通过调参确定政协提案中标题和正文的权重比；将词向量加权平均的结果做为提案向量。

本文采用的提案向量化方法既考虑了词与词之间的语义信息，也考虑了政协提案的结构信息，因此这种方法实现了提案信息的深入挖掘，向量化效果更好。

(2) 政协提案的向量化算法实现

政协提案的向量化过程包括三个步骤：训练词向量、特征选择、计算标题和正文的权重比。

1) 训练词向量

本文使用开源的 Word2vec 工具中的 Skip-Gram 模型对预处理后的政协提案中的词训练词向量。

使用 Word2vec 训练词向量之前涉及三个重要的参数选择，分别是窗口大小、向量维度、参与训练的词的最低词频。对于窗口大小，本文选择经验值 5；向量维度的取值一般在 50-300 之间，为了使词向量能够表达更多的文本语义信息，本文选择向量维度为 300；由于提案中可能存在出现次数很少的生僻词，这些词一般情况下不会影响文本的语义，因此本文将参与训练的词的最低词频设置为 2，表示在

所有提案中出现总次数低于 2 的词不参与训练。将参数设置好后开始训练词向量，然后使用训练好的词向量表示提案向量。

2) 特征选择

经过预处理的政协提案数据虽然去除了停用词，但直接将每件提案中所有词的词向量求均值作为提案向量会存在两个问题：忽略了不同词对提案的重要性不同；提案中不重要的词参与提案的向量化表示会对向量表示引入噪声。

以上两个问题都会影响提案的向量化表示效果，因此本文对提案中的词进行特征选择。常用的特征选择方法有 TF-IDF、卡方检验等。

卡方检验适合对短文本数据进行特征选择，在长文本数据的特征选择中容易造成低频词缺陷，而 TF-IDF 适合长文本数据的特征选择。由于政协提案均为长文本数据，因此本文使用 TF-IDF 算法计算提案中每个词的 TF-IDF 值并选择最优的 TF-IDF 阈值，使用大于 TF-IDF 阈值的词向量计算提案向量。

3) 计算标题和正文的权重比

标题能在很大程度上反映文章的主旨内容，因此标题和正文对文章的重要性不同[50]。政协提案数据也由标题和正文构成，因此本文在计算提案向量时对标题赋予更高的权重。

综合考虑以上因素，本文计算政协提案向量的公式如下：

$$V_{提案}=\frac{V_{标题}*(TF\cdot IDF)*W_{标题}+V_{正文}*(TF\cdot IDF)}{N} \tag{4-1}$$

其中 $V_{提案}$表示政协提案向量，$V_{标题}$表示政协提案的标题中的词向量，$W_{标题}$表示为标题中的词赋予的权重，$V_{正文}$表示政协提案的正文中的词向量，N 表示每件提案中的单词总数。

由于无法直接衡量 TF-IDF 阈值、标题-正文权重比的不同取值对提案向量的影响，考虑到提案向量的优劣会影响聚类效果，聚类的类别数 k 也会影响聚类效果，因此本文通过调节 TF-IDF 阈值、标题-正文的权重比、聚类的类别数 k，观察聚类效果的评价指标的变化，得到使聚类效果的评价指标最好的参数值组合作为最优参数值组合。

4.2.2 K-means 聚类

本文在 4.2.1 节中介绍了政协提案的向量化算法，将每件提案映射成了空间中的一个点。本节使用 K-means 聚类算法对向量化后的提案聚类，使聚类形成的每一类中的提案的主题相同。在本节使用的 K-means 聚类算法中，点之间的距离定

义为欧式距离。

(1) 聚类评价指标的选择

K-means 聚类表示对向量化后的提案数据聚类，使每一类中的提案具有相同的主题，其核心在于 k 值的选取。由于对提案数据没有足够的先验知识，不能从主观上确定 k 值的大小，因此需要选择合适的聚类效果评价指标，通过调节 k 值的大小来观察聚类评价指标的变化，最终确定合适的 k 值。常用的聚类评价指标有轮廓系数、平均半径、平均直径等，本文选择使用最广泛的轮廓系数作为评价指标。

(2) 直接遍历所有参数存在计算量大的问题

如上节所述，提案向量化过程中的 TF-IDF 阈值、标题-正文的权重比的不同取值会影响提案向量化效果，从而影响聚类效果；本节介绍的 K-means 聚类算法的 k 值的不同取值也会影响聚类效果。为了得到最优的聚类效果，最直接的想法是在一定范围内遍历这些参数的不同组合，计算每种参数组合的轮廓系数，选择使轮廓系数最大的参数组合作为最优的参数组合。

然而，这种方法会导致巨大的计算量。论据如下：

为了确定在提案向量化过程中的 TF-IDF 阈值、标题-正文的权重比、k 值这三个参数的大小，本文对这三个参数在一定范围内进行遍历，每个参数的遍历取值范围及步长如下表 4-5 所示：

表 4-5 结巴分词的词性标注结果

Table 4-5 Results after removing the stop word

参数	取值范围	步长
TF-IDF 阈值	0～0.15	0.01
标题-正文权重比	1～10	1
k 值	3～30	1

从表 4-5 中可以计算出三个参数共有 4480(16*10*28)种不同的组合方式，如果直接遍历所有的参数组合，从中选择使轮廓系数最大的参数组合作为最终的参数，将会耗费大量的计算资源和时间。在本文实验所用的笔记本电脑上，程序运行一次需要 3 分钟，所以理论上遍历所有的参数组合需要运行 9 天，因此会耗费大量的计算资源和时间。

(3) 通过迭代调参解决计算量大的问题

本文采用迭代调参的方式解决计算量大的问题。具体方法是：给定每个参数的初始值，控制两个参数不变，遍历另一个参数，绘制聚类的轮廓系数随该参数的变化曲线，选择使轮廓系数最大的参数值，更新该参数。采用控制变量法循环遍历每个参数并不断更新参数值，直到三个参数稳定。经过 30 轮迭代调参后，三个参数

达到稳定状态。采用这种方法，总共遍历了 1620((28+16+10)*30)种不同的参数组合方式，同样配置的情况下，只用了直接遍历三分之一的时间。

稳定后的参数值分别如下表 4-6 所示：

表 4-6 提案表示和聚类的最优参数

Table 4-6 Proposal representation and clustering optimal parameters

参数	最终结果
TF-IDF 阈值	0.04
标题-正文权重比	5:1
k 值	8

因此本文最终在 TF-IDF 阈值为 0.04、标题-正文权重比为 5:1 的情况下计算政协提案向量，并将政协提案聚为 8 类。每一类中的提案属于同一个主题，每个主题中的提案数如下表 4-7 所示：

表 4-7 聚类结果中每一类的提案数

Table 4-7 Number of proposals for each category in the clustering results

主题	提案数
主题 1	68
主题 2	78
主题 3	3
主题 4	432
主题 5	85
主题 6	57
主题 7	60
主题 8	15

由表 4-7 可知，本文没有对主题定义主题名来描述每个主题，在下节的关键词提取中，本文从每个主题中提取出三个关键词作为对每个主题的主旨描述。

4.3 关键词提取模块的关键词提取方法

关键词提取是指从每个主题的多件政协提案中提取关键词，使提取出的关键词可以准确的表示每个主题的主旨内容，下面介绍两种关键词提取算法。

4.3.1 简单提取法

在 4.2 节我们提出了一种政协提案向量化算法，并使用 K-means 聚类算法对向量化后的政协提案进行了聚类，使每一类中的提案属于同一个主题。例如本文将2018年北京市的政协提案聚成了8类。本节对每个主题下的多件提案提取关键词。

从每个主题的多件提案中提取关键词的核心是计算每个主题中的每个词的权重，将权重最大的前几个词作为关键词，提取的关键词的数量可以自行设置。对每个主题，本文将每个词在主题中的 TF-IDF 值作为权重值，并对主题中的词按权重值排序，选择权重最大的前三个词作为关键词。

对每个主题，每个词在主题中的 TF-IDF 值的计算方法和关键词提取方法如下所述：

将每个主题中同一个词的 TF-IDF 值相加作为每个词在每个主题中的权重，对每个主题中的词按权重排序，提取每个主题中权重最大的前三个词作为关键词。

本文对 2018 年北京市的政协提案聚类形成的 8 个主题提取关键词，结果如下表 4-8 所示：

表 4-8 每个主题中的关键词提取结果

Table 4-8 Keyword extraction results for each topic

主题	关键词
主题 1	文化，保护，艺术
主题 2	产业，企业，创新
主题 3	污泥，资源化，处理
主题 4	建设，城市，发展
主题 5	服务，社区，医疗
主题 6	停车，共享单车，老年代步车
主题 7	教育，教师，学生
主题 8	垃圾，垃圾分类，回收

由表 4-8 可知，从每个主题中提取出的关键词大部分长度较短(以下简称“短词”)，词语的语义表达能力有限，使用这些关键词从微博中采集数据可能会存在大量与主题无关的微博数据。因此本文对上述关键词提取算法进行改进，以期得到语义表达能力更强的词。

4.3.2 简单提取法的改进

由上文可知，简单提取法提取出的关键词长度较短，语义表达能力有限，而更

长的词的语义表达能力则更强。例如，“垃圾分类”比“垃圾”表达的含义更为清楚。如果能从每个主题中提取出更长的词，则能更清楚的表达每个主题的主旨内容。

通过阅读文献，发现如果某个词前后词的词性属于[‘n’,‘nr’,‘ns’,‘nt’,‘nz’,‘vn’,‘an’,‘f’]，则该词可与其前后词构成词组[51]。本文借鉴这个语言规则发现提案中可能的“长词”。具体的实现过程如下：

在提案不去除停用词的情况下，使用下面的步骤从每个主题中提取“长词”：

(1) 提取候选关键词。将聚类得到的每个主题中同一个词的 TF-IDF 值相加作为每个词在每个主题中的权重，对每个主题中的词按权重排序，提取权重最大的三倍于提案数量的词作为每个主题的候选关键词。

(2) 发现候选“长词”。循环遍历从每个主题中提取出的候选关键词，找到每个候选关键词在每个主题的提案原文中的位置，判断每个候选关键词在提案原文中的前一个词的词性是否属于[‘n’,‘nr’,‘ns’,‘nt’,‘nz’,‘vn’,‘an’,‘f’]，若是，则将候选关键词的前一个词与候选关键词组成新的关键词；同理可将候选关键词与其后一个词组成新的关键词；若候选关键词在提案原文中的前一个词和后一个词的词性均属于[‘n’,‘nr’,‘ns’,‘nt’,‘nz’,‘vn’,‘an’,‘f’]，则将候选关键词与其前后各一个词共同构成新的关键词。

通过以上两个步骤，在提案不去除停用词的情况下，发现了每个主题中的一组候选“长词”。

同理，在提案去除停用词的情况下，重复上述过程，发现了每个主题中的另一组候选“长词”。

提取最终的“长词”。将提案不去除停用词和提案去除停用词两种情况下，根据语言规则得到的两组候选“长词”取交集得到每个主题最终的候选“长词”，再从每个主题最终的候选“长词”中选择权重最大的前三个词作为最终的“长词”。因此，最终从每个主题中提取的“长词”如下表 4-9 所示：

表 4-9 算法改进后每个主题中的关键词提取结果

Table 4-9 Keyword extraction results for each topic after the algorithm is improved

主题	关键词
主题 1	文化遗产，文化中心，传统文化
主题 2	知识产权，科技创新，科技创新中心
主题 3	污泥利用，污水处理，污泥处理
主题 4	轨道交通，美丽乡村，基础设施
主题 5	分级诊疗，医疗机构，养老服务
主题 6	共享单车，老年代步车，物业管理
主题 7	中小学生，学前教育，融合教育
主题 8	垃圾分类，再生资源，建筑垃圾

由表 4-9 可知，改进后的关键词提取算法从每个主题中提取的关键词长度较长，因而具有更强的语义表达能力。

本文将在第五章对两种关键词提取算法提取出的关键词进行对比实验，判断哪种算法提取的关键词更有效。

4.4 情感分类模块的微博数据预处理

本文对微博数据的预处理步骤包括数据清洗、分词和停用词过滤。由于对微博数据的分词和停用词过滤与对政协提案的分词和停用词过滤处理方法相同，因此本节不再详述，只介绍对微博数据的数据清洗过程。

通过观察采集的微博数据，发现微博数据中存在一些干扰字符和无效微博，主要包括：

(1) 从微博中爬取的数据会存在 HTML 符号，如<p>，
等干扰字符，这些符号都属于噪声，需要从数据中删除；

(2) 采集的微博数据中会存在表情符、标点符单独构成的微博和采集失败的空微博，这些微博会使情感分类模型出错，需要删除。

数据清洗的目的是从原始数据中删除以上干扰字符和无效微博，从而筛选出纯文本数据用于情感分类。针对以上问题，本文采用不同的解决方法：

(1) 在 Python 程序中使用正则表达式删除微博数据中的干扰字符；

(2) 在 Excel 表格中对原始微博数据排序，使同类型的微博数据在表格中相邻，如所有的表情符相邻，空微博相邻，然后手动删除所有无效微博。

将清洗后的微博数据重新保存到新的 Excel 表格中，作为后续情感分类模型的训练、测试和预测数据。

4.5 情感分类模块的情感分类模型搭建

4.5.1 情感分类模型的流程图

本文设计了基于双向 LSTM 的情感分类模型，模型流程图如图 4-2 所示，从图中可以看出，基于双向 LSTM 的情感分类模型是一个五层结构的模型，以下对每一层分别展开介绍。

第一层为输入层，输入层负责将训练数据输入双向 LSTM 情感分类模型中。对输入数据的格式说明如下：由于文本数据不能直接输入到情感分类模型中，因此需要将词语进行数字化，常用的方法是将词语使用 Word2Vec 向量化，向量维度取

50-300，但由于微博数据量较大，将使用 Word2Vec 向量化后的微博数据输入情感分类模型中会占用大量内存。因此本文将每个词语映射为一个数字，则句子映射为一个数字序列，计算所有微博数据的平均长度，将每个数字序列化后的句子统一为平均长度，具体为将大于平均长度的微博数据截断，小于平均长度的微博数据补 0，将使用数字序列化并统一长度后的句子输入情感分类模型中进行后续处理。

第二层为 Embedding 层，Embedding 层的作用与 Word2Vec 相同，Embedding 层基于句子中的语义信息将每个数字化后的词转化为固定长度的向量。

第三层为双向 LSTM 层，是情感分类模型的核心层，作用是分别对每个句子中的词按照从前向后的顺序和从后向前的顺序进行处理，提取每个句子中的上下文信息，然后将两种方式提取的信息进行组合，作为下一层的输入信息。为了防止过拟合，本文使用了 Dropout 和正则化。

第四层为全连接层，主要对双向 LSTM 层输入进来的数据进行特征提取，从而实现数据降维。

第五层为激活函数层，作用是对全连接层输入进来的数据进行非线性化处理，从而引入非线性特征，提高模型的表达能力。

第六层为输出层，如果是二分类模型，则使用 Sigmoid 函数输出两个类别；如果是多分类模型，则使用 Softmax 函数输出多个类别。本文的情感分类模型为二分类模型，因此在输出层使用 Sigmoid 函数输出 0，1 值。

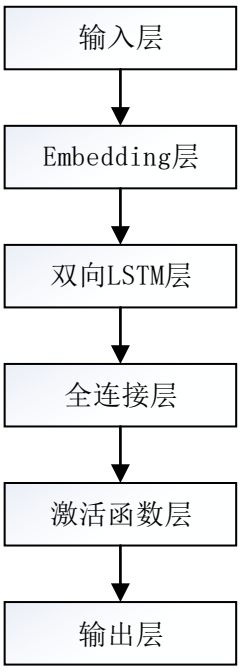


图 4-2 情感分类模型流程图

Figure 4-2 Sentiment classification model flow chart

4.5.2 情感分类模型的实现

本节给出情感分类模型的实现程序，程序的核心部分如图 4-3 所示，可以看出，情感分类模型由五部分组成，分别是：从深度学习框架 Keras 库中导入相应的模型结构；使用 train_test_split()函数将输入数据划分为训练集、验证集和测试集，三者的比例可以自行设置，本文设置为 8:1:1；定义模型中的参数值大小；设计模型结构；使用训练集中的数据训练模型并使用测试集中的数据进行准确率测试。

```
from keras.layers import Activation, Dense,Embedding,LSTM,Bidirectional
from keras.models import Sequential        #序贯模型
from keras.preprocessing import sequence   #sequence.pad_sequences,用于统一句子长度
from sklearn.model_selection import train_test_split  #划分数据集
from keras import regularizers

def model(X,y,MAX_SENTENCE_LENGTH,vocab_size):
    epochs_list=[]
    Xtrain, Xtest_val, ytrain, ytest_val = train_test_split(X[:8883], y[:8883], test_size=0.2, random_state=42)# 数据划分
    Xval, Xtest, yval, ytest = train_test_split(Xtest_val, ytest_val, test_size=0.5, random_state=42)  # 数据划分
    EMBEDDING_SIZE = 100      #词向量维度
    HIDDEN_LAYER_SIZE = 32    #LSTM中隐藏层的维度
    BATCH_SIZE = 300          #每次从训练集中选择300个样本训练并更新参数
    NUM_EPOCHS = 15           #训练轮数
    for i in range(NUM_EPOCHS):
        epochs_list.append(i+1)
    model = Sequential()
    model.add(Embedding(vocab_size, EMBEDDING_SIZE,input_length=MAX_SENTENCE_LENGTH))
    model.add(Bidirectional(LSTM(HIDDEN_LAYER_SIZE, dropout=0.8, recurrent_dropout=0.8,
                                 kernel_regularizer=regularizers.l2(0.001),recurrent_regularizer=regularizers.l2(0.001)
                                 bias_regularizer=regularizers.l2(0.001))))
    model.add(Dense(1))
    model.add(Activation("sigmoid"))
    model.compile(loss="binary_crossentropy", optimizer="adam",metrics=["accuracy"])
    history=model.fit(Xtrain, ytrain, batch_size=BATCH_SIZE, epochs=NUM_EPOCHS,validation_data=(Xval, yval)) # 模型训练
    score, acc = model.evaluate(Xtest, ytest)  # 预测
    return score, acc,epochs_list,history,model,Xtest,ytest,Xtrain,ytrain,Xval,yval
```

图 4-3 情感分类模型的实现程序

Figure 4-3 Implementation program of sentiment classification model

4.6 本章小节

本章前三节主要介绍了对政协提案划分主题并提取关键词的方法，包括对政协提案进行数据预处理的过程中遇到的问题及解决方法、政协提案的向量化表示方法和聚类方法、两种从聚类结果中提取关键词的方法。第四节和第五节主要介绍了情感分类模型的设计方法和实现步骤，包括情感分类模型的输入数据的预处理方法和情感分类模型的搭建。

5 实验结果分析

本章介绍本文的实验结果并对结果进行分析，主要分为五部分。第一部分介绍本文实验过程中依托的硬件配置和软件环境；第二部分介绍本文提取的两组关键词的对比实验的设计方法，并对结果进行分析说明；第三部分介绍使用微博爬虫程序采集每个关键词的微博数据的过程中存在的问题和解决方法，并展示每个关键词的采集结果；第四部分介绍相关舆情的分析结果并对结果进行分析说明；第五部分为本章小结。

5.1 实验环境

(1) 处理器：2.3 GHz Intel Core i5
(2) 内存：8 GB 2133 MHz LPDDR3
(3) 开发环境：Anaconda3 中的 Jupyter Notebook，Vscode
(4) 开发语言：Python
(5) 主要的工具包：Scikit-Learn，Keras，Gensim，Scrapy，Selenium 等

5.2 关键词有效性分析

本文在第四章中详细介绍了两种关键词提取算法，并分别使用每种算法从每个主题中提取出三个关键词，因此对于每个主题提取出两组关键词，分别简称“短词”和“长词”。本节综合考虑实验目的和可行性，设计了有效的关键词对比试验，通过分析实验结果对比两组关键词的优劣，选择质量高的关键词从微博中采集数据，用于进行舆情统计分析。

5.2.1 关键词质量评价的方法

通常将提取出的关键词与专家标注的关键词对比，计算准确率、召回率和 F 值，作为关键词提取质量的评价标准。然而由于本文研究的政协提案没有专家标注的关键词，而人工标注关键词存在很大的主观性，因此本文通过计算每个主题对应的关键词获取的微博数据与该主题中的提案之间的相似性，评价关键词优劣。

本文采用如下方法评价两组关键词的质量：

(1) 使用微博爬虫程序分别采集与“长词”和“短词”相关的微博数据。

(2) 对于每个主题，从“长词”采集的微博数据中随机选择部分微博数据构成“微博文档”，使“微博文档”的长度近似等于每个主题中政协提案的平均长度，因此对于本文的 8 个主题，使用“长词”采集的微博数据共构成 8 篇“微博文档”；同理，对“短词”采集的微博数据重复相同的步骤，也构成 8 篇“微博文档”。

(3) 对每个主题，计算“微博文档”和政协提案之间的欧式距离，用这个距离衡量“长词”和“短词”的提取质量。首先把所有的政协提案和“微博文档”转换为向量，转换方法在下面的 5.2.2 节描述。然后对每个主题，计算“长词”构成的“微博文档”与主题中的政协提案之间的欧式距离并取均值，简称为“长词距离”；同理，计算“短词”构成的“微博文档”与主题中的政协提案之间的欧式距离并取均值，简称为“短词距离”；比较“长词距离”与“短词距离”的大小，距离越小，说明关键词采集的微博数据的内容与政协提案的内容越相似，关键词越能反映提案主旨。

这里构造“微博文档”的原因如下：由于政协提案数据均为长文本，而微博数据均为短文本，直接计算每个主题下的每条微博数据与对应主题中的政协提案数据的相似性不仅计算量大，而且将长文本与短文本直接进行相似性计算误差较大。因此本文计算每个主题中的政协提案的平均长度，从每个主题对应的微博数据中随机选择部分数据进行组合，使组合后的微博文本长度近似等于该主题中政协提案的平均长度，将组合后的微博文本数据称为“微博文档”。因此使用两组关键词采集的微博数据共构造出 16 篇“微博文档”。

5.2.2 关键词质量评价的具体实现

对每个关键词爬取 2018 年的所有微博数据用于质量评价会耗费大量时间和精力。由于 2018 年北京市的政协提案在 1 月底提出，因此使用每个关键词在 2018 年 2 月-2018 年 12 月的微博数据构造“微博文档”。然而“长词”和“短词”各有 24 个，两组关键词共有 48 个，如果对每个关键词都从微博中采集 2018 年 2 月-2018 年 12 月的所有微博数据会耗费大量时间和精力。以每个关键词在 2018 年平均有一万条微博计算，使用本文设计的微博爬虫程序采集微博数据，需要 9 个小时，则 24 个关键词需要花费 9 天时间。

对于每个关键词，本文首先在 2018 年 2 月-2018 年 12 月的时间区间内随机取样，然后在这些取样时间区间内对该关键词进行微博爬取，这样就可以减少微博爬取的时间。具体如下：为了使爬取的微博数据均匀分布在 2018 年 2 月-2018 年 12 月之间，每个关键词在每个月都要爬取部分数据；为了使爬取的数据具有更大的随机性，本文爬取每个月不同日期的数据。如爬取 2018 年 4 月 7 日-2018 年 4 月 10

日、2018 年 5 月 11 日-2018 年 5 月 14 日的数据等。将两组关键词爬取的微博数据按关键词保存到 Excel 表格中。使用 4.4 节介绍的数据清洗方法对所有的微博数据进行数据清洗，去除干扰字符和无效微博，重新保存到新的 Excel 表格中。

下面介绍“微博文档”和政协提案的向量化表示的具体实现与相似性计算。

(1) “微博文档”和政协提案的向量化表示

使用 4.1 节介绍的数据预处理方法对“微博文档”和政协提案进行分词、去除停用词，使用 Word2Vec 算法对预处理后的“微博文档”和政协提案中的词进行向量化训练，得到每个词的词向量，并使用 TF-IDF 算法计算每个词的权重，最后使用每个词的词向量和 TF-IDF 值将“微博文档”和政协提案进行向量化表示，向量化的公式如下：

$$V_{文档/提案} = \frac{V_{词} * (TF \cdot IDF)}{N} \tag{5-1}$$

其中 $V_{文档/提案}$表示“微博文档”或政协提案向量，$V_{词}$表示“微博文档”或政协提案中的每个词的词向量，$TF \cdot IDF$表示每个词所对应的权重，N表示“微博文档”或政协提案中的词数。

(2) “微博文档”和政协提案的相似性计算

对每个主题，将由“长词”采集的微博数据构造的“微博文档”和由“短词”采集的微博数据构造的“微博文档”分别与主题中的每件政协提案进行相似性计算，本文使用欧式距离计算相似性，欧式距离越小，表示相似性越大，相似性的计算公式为：

$$D_k = \frac{\sum_1^{N_k} d_i}{N_k} \ , k = 1,2,\cdots 8 \tag{5-2}$$

其中D_k表示“微博文档”与主题中的每件提案的平均欧式距离，d_i表示“微博文档”与主题中的每件提案之间的欧式距离，N_k表示每个主题中的提案数量。

将从每个主题中提取出的两组关键词采集的微博数据分别构成的“微博文档”与每个主题中的提案按公式(5-2)计算平均欧式距离，结果如图 5-1 所示。图中的横坐标表示主题，纵坐标表示欧式距离。蓝色柱状图表示“短词”对应的“微博文档”与相应主题中的提案之间的平均欧式距离，红色柱状图表示“长词”对应的“微博文档”与相应主题中的提案之间的平均欧式距离。

从图中可以看出，对每个主题，“长词”对应的“微博文档”与主题中的政协提案之间的平均欧式距离更小，因此由“长词”采集的微博数据与每个主题中的提案内容更相似，“长词”具有更强的表达能力，更能表示每个主题的主旨内容。因此本文使用“长词”从微博中采集数据进行舆情分析。

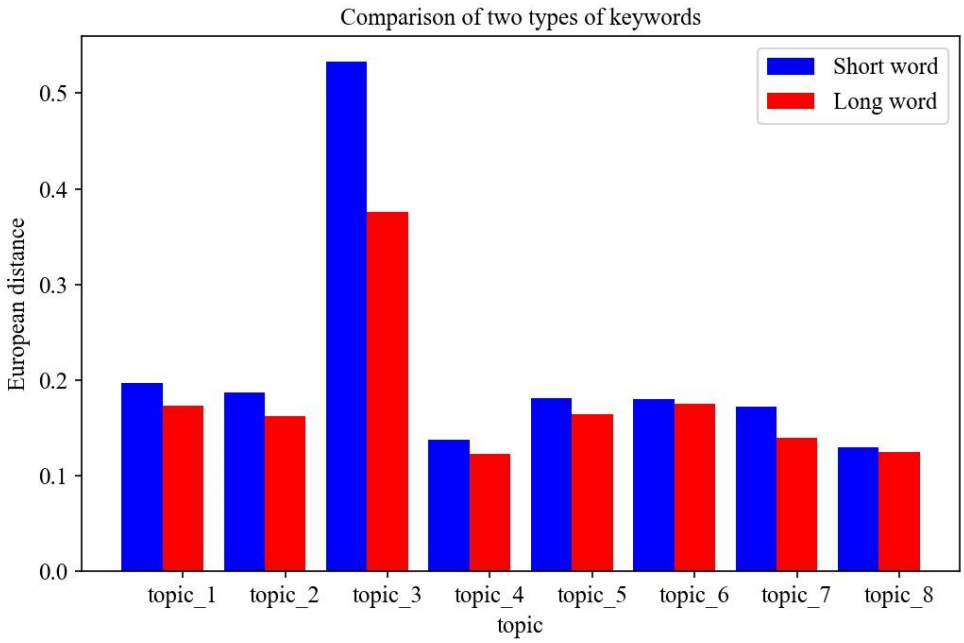


图 5-1 关键词有效性分析实验结果

Figure 5-1 Key words effectiveness analysis experimental results

5.3 微博数据采集与处理

虽然 2018 年北京市的政协提案在 1 月底提出，民众对新提出的政协提案的微博评论从 1 月底开始，但政协提案提出前微博中也存在各方面的讨论。本文由于情感分类模型的需要，采集 2018 年 1 月-2018 年 12 月的微博数据。

由于微博每次最多只能返回 50 页数据，为了保证数据采集的完整性，使用每个关键词采集微博数据时，首先查看给定时间段内的微博数量是否小于 50 页，若小于 50 页，则采集该时间段内的微博数据，否则缩小时间段，不断重复该过程，直到完成每个关键词在 2018 年的微博数据采集。

将每个关键词采集的微博数据保存到 Excel 表格中，使用 4.4 节介绍的数据清洗方法对微博数据进行数据清洗，重新保存到新的 Excel 表格中。本文对 8 个主题下的 24 个关键词分别采集 2018 年的微博数据，共采集了 223963 条数据。经过数据清洗后可用的微博数据有 169803 条，其中所有关键词在 2018 年 1 月的微博数据共有 17494 条，本文对这些数据进行人工打标，作为情感分类模型的训练数据。所有关键词在 2018 年 2 月-2018 年 12 月的微博数据共有 152309 条，本文使用这些数据分析人们对每个主题的关注度大小和演进趋势、情感倾向和演进趋势。

为了更具体地介绍每个关键词爬取的微博数量，在表 5-1 中列出了 2018 年北

京市政协提案提出后(即 2018 年 2 月后)网民发布的微博数量情况。

表 5-1 每个主题和关键词的微博数量
Table 5-1 Number of Weibos each topic and keyword

关键词	微博数
文化遗产	7074
文化中心	9589
传统文化	14318
知识产权	3940
科技创新	15252
科技创新中心	3852
污泥利用	2768
污水处理	3174
污泥处理	1774
轨道交通	9765
美丽乡村	2342
基础设施	15077
分级诊疗	9668
医疗机构	6671
养老服务	2821
共享单车	17209
老年代步车	2309
物业管理	4459
中小学生	12806
学前教育	1930
融合教育	778
垃圾分类	2694
再生资源	645
建筑垃圾	1394

5.4 舆情分析实验设计与结果分析

本文对政协提案进行相关舆情统计分析，主要从两方面展开分析，分别是分析在政协提案提出后人们对每个主题的关注度大小和关注度演进趋势变化、情感总

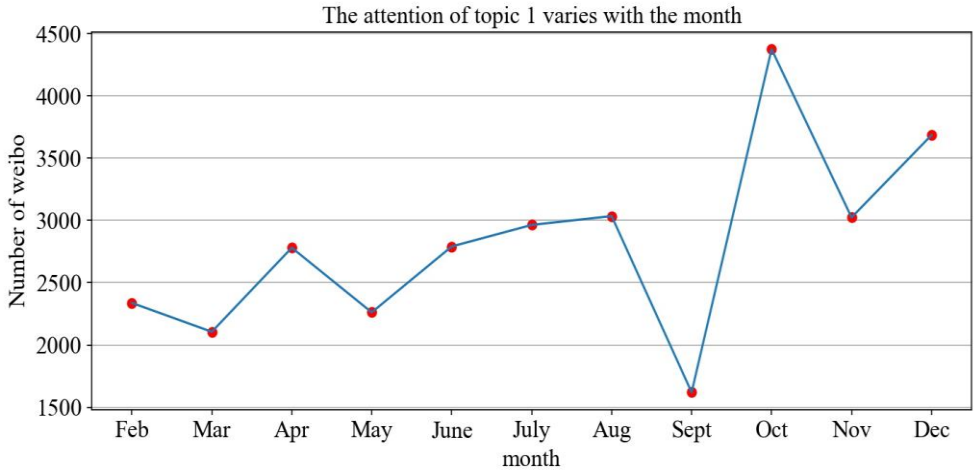
体倾向和情感演进趋势变化。

5.4.1 关注度分析

关注度分析分为两个方面，分别是关注度演进趋势分析和关注度大小分析。关注度演进趋势分析通过统计每个主题在不同时间段的微博数量，从而绘制微博数量随时间的变化曲线进行分析。关注度大小分析指在政协提案提出后，对每个主题在 2018 年 2 月-2018 年 12 月的微博数量和每个主题中的提案数量进行对比分析。

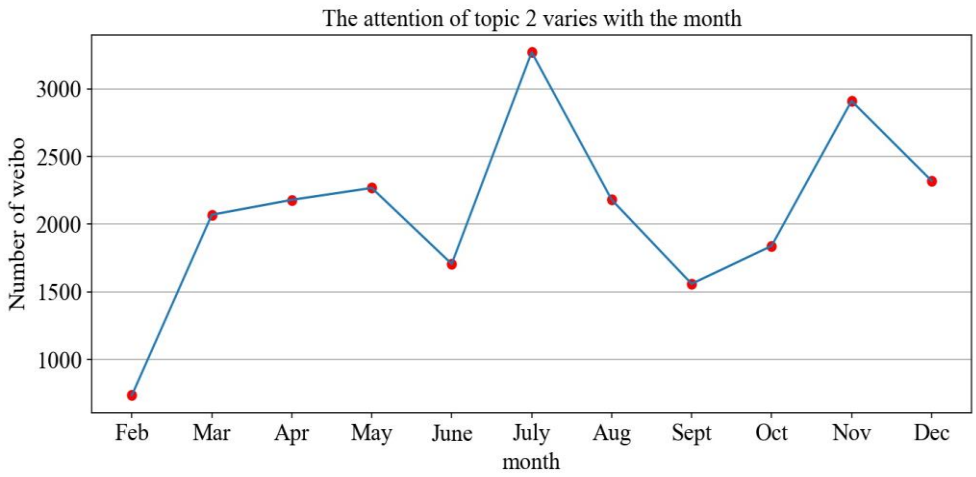
(1) 关注度演进趋势分析

本文以月为单位统计每个主题在不同月份的微博数量，并绘制每个主题在不同月份的微博数量随月份的变化曲线，结果如图 5-2 所示：



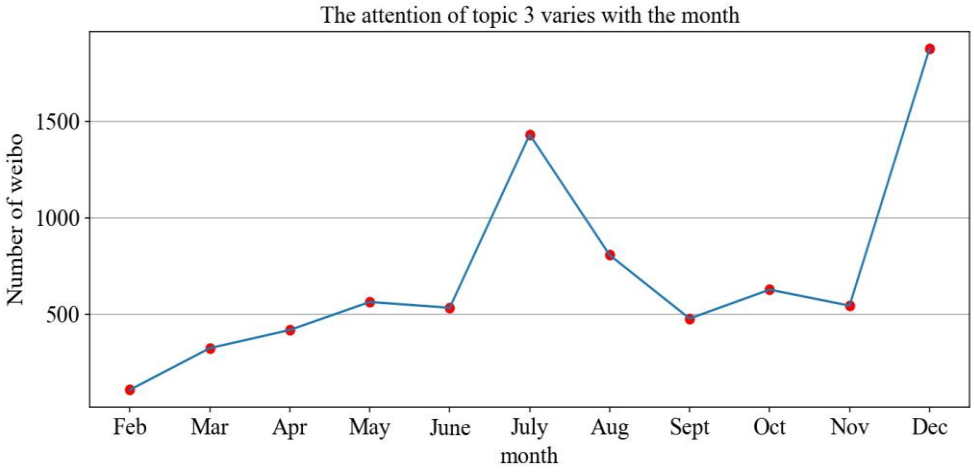
a) 主题 1 的关注度变化曲线

a) Attention curve of topic one

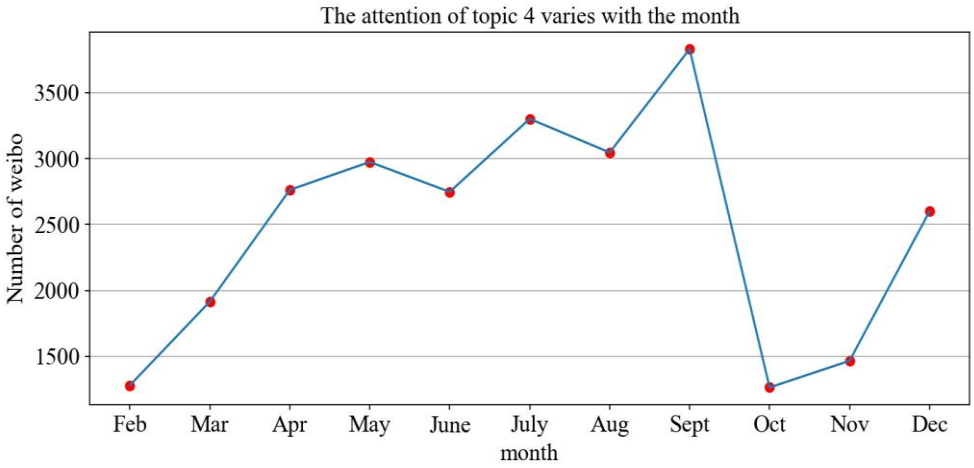


b) 主题 2 的关注度变化曲线

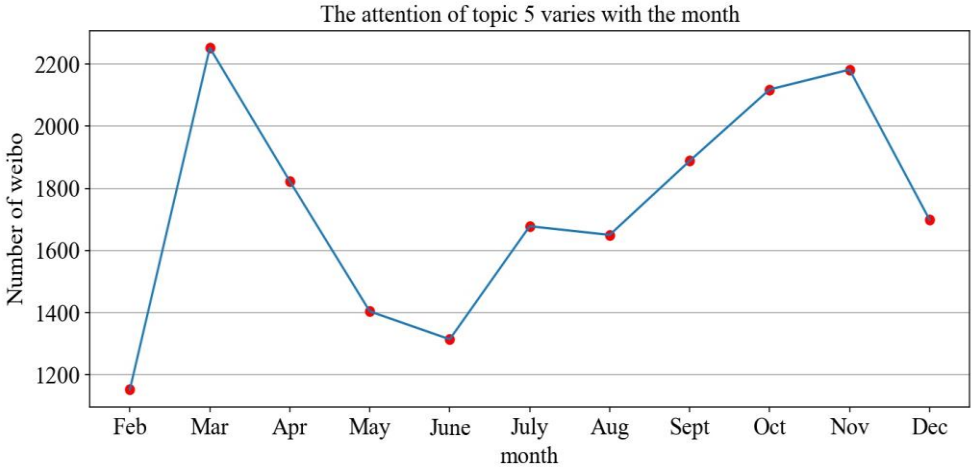
b) Attention curve of topic two



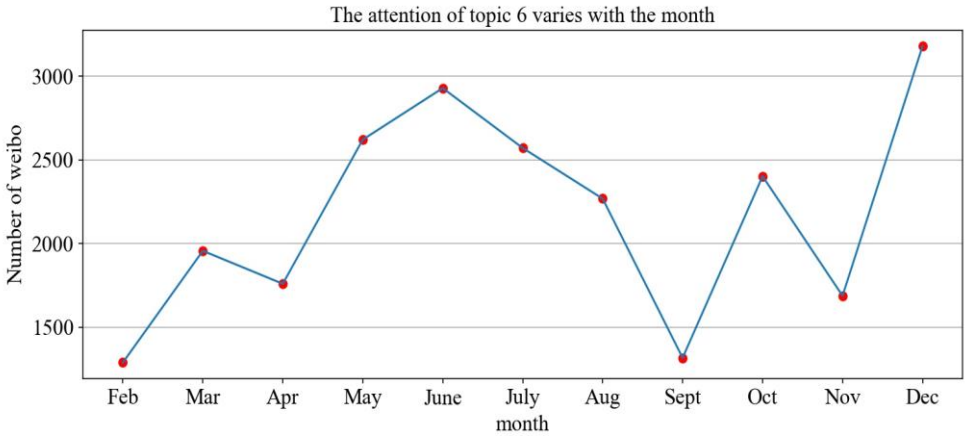
c) 主题 3 的关注度变化曲线
c) Attention curve of topic three



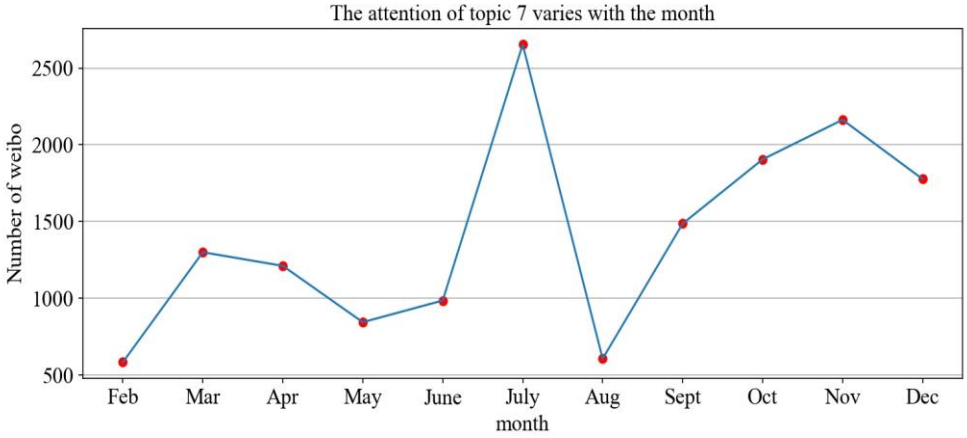
d) 主题 4 的关注度变化曲线
d) Attention curve of topic four



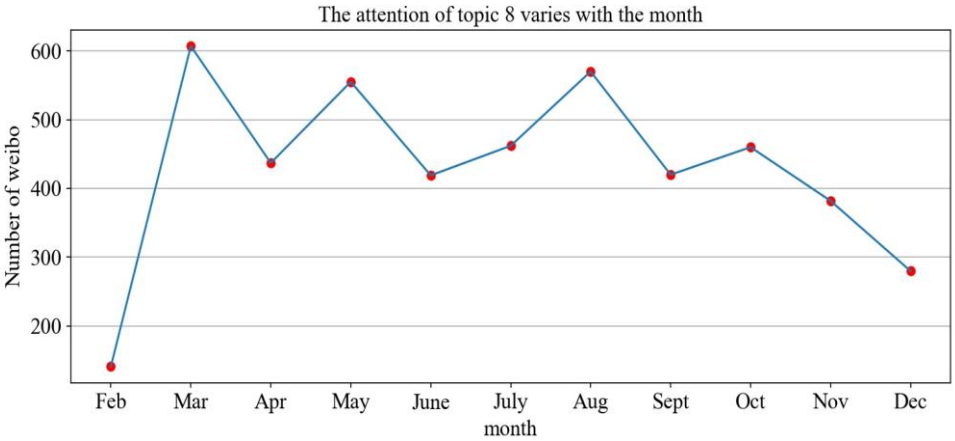
e) 主题 5 的关注度变化曲线
e) Attention curve of topic five



f) 主题 6 的关注度变化曲线
f) Attention curve of topic six



g) 主题 7 的关注度变化曲线
g) Attention curve of topic seven



h) 主题 8 的关注度变化曲线
h) Attention curve of topic eight

图 5-2 民众对每个主题的关注度变化曲线
Figure 5-2 People's attention curve for each topic

图 5-2 表示每个主题对应的每个月的微博数量随月份的变化曲线，其中横坐标表示月份，纵坐标表示微博数量。从图 5-2 中可以得出以下几个结论：

1) 在政协提案提出后，人们对每个主题的关注度虽然不断起伏变化，但整体都呈现上升趋势，由此可知政协提案引起了人们的关注，具有一定的社会影响力；

2) 人们对每个主题的关注度都存在峰值，而峰值出现的时间不同。比如：主题 5 和主题 8 的峰值出现在三月，说明这两类主题的政协提案一经提出就受到了人们的较大关注；主题 3 和主题 6 在 2018 年的峰值出现在 12 月，说明这两类主题的政协提案刚提出没有受到太多关注，但之后的影响力在网络中迅速传播；

3) 总体上，人们在 2 月对每个主题的关注度都很低，之后对每个主题的关注度不断起伏变化，但关注度基本不会低于对 2 月的关注度。不过主题 1 例外，人们对主题 1 在 9 月的关注度远远低于 2 月。

(2) 关注度大小分析

本文统计每个主题在 2018 年 2 月-2018 年 12 月的微博数量，将每个主题对应的微博数量与提案数量进行对比，分析民众和政协委员对每个主题的关注度异同。将每个主题对应的微博数量和提案数量绘制成柱状图，如图 5-3 所示：

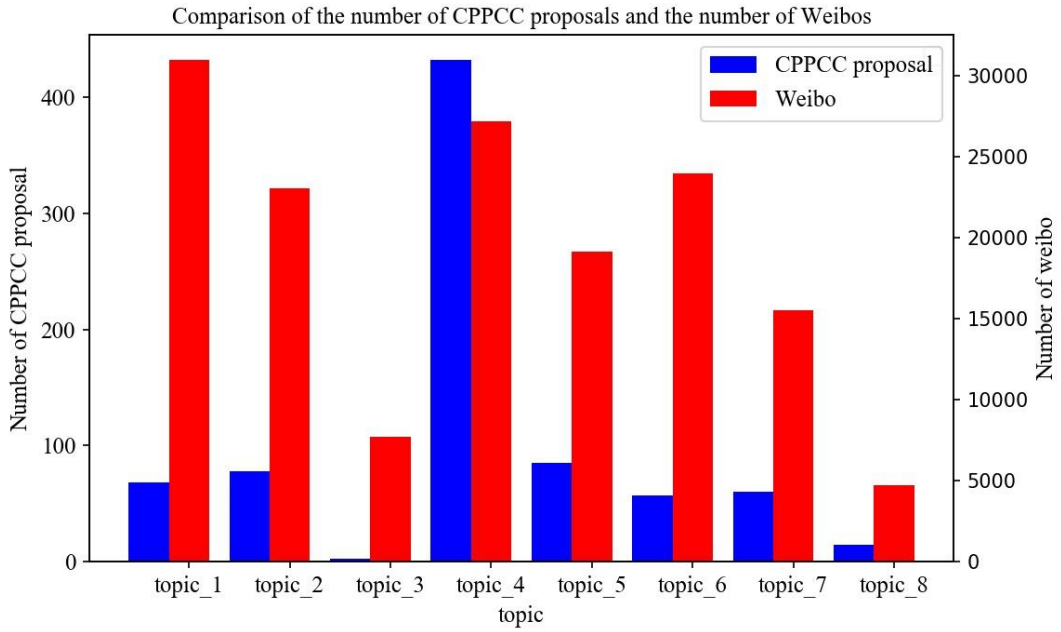


图 5-3 政协委员和民众对每个主题的关注度

Figure 5-3 The attention of CPPCC members and people on each topic

在图 5-3 中将不同主题对应的微博数量和提案数量进行对比，图中设置了两个纵坐标系，左边的纵坐标表示政协提案数，右边的纵坐标表示微博数，横坐标表示主题，蓝色柱状图表示每个主题的提案数量，红色柱状图表示每个主题对应的微博

数量。

通过分析图 5-3，可以得出以下结论：

1) 主题 4 的政协提案数和微博数都很多，说明主题 4 受到政协委员和民众共同的广泛关注，同时说明主题 4 中的提案贴近民意，且具有广泛的社会影响力。

2) 主题 1、主题 2、主题 5、主题 6、主题 7 的政协提案数量相差不大，且每个主题中的提案数都不多，但是这几个主题对应的微博数都很多，尤其是主题 1 的微博数更是超过了主题 4 的微博数。因此虽然政协委员对这几个主题的关注程度一般，但民众对这几个主题的关注程度较高。

3) 主题 3 和主题 8 的政协提案数和微博数都很少，说明政协委员和民众对这两个主题的关注程度都很低。

5.4.2 情感分析

情感分析主要从两方面展开：情感演进趋势分析和情感倾向总体分析。情感演进趋势分析通过统计每个主题在不同时间段的不同情感极性的微博数量，从而绘制不同情感极性的微博数量随时间的变化曲线进行分析。情感倾向总体分析指在政协提案提出后，对每个主题在 2018 年 2 月-2018 年 12 月的不同情感极性的微博数量进行统计，绘制每个主题不同情感极性的微博数量柱状图。

(1) 情感分类模型训练

由于情感分析需要知道每条微博数据的极性，人工对所有的微博数据标注极性显然不现实，因此本文对少量数据人工标注情感极性，使用标注数据训练模型，将训练好的模型用于未标注数据的情感极性预测。模型训练步骤如图 5-4 所示：

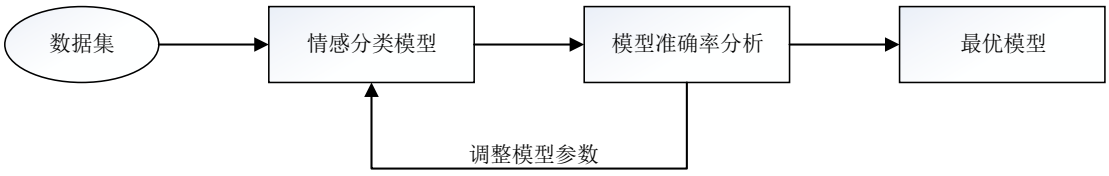


图 5-4 情感分类模型训练流程图

Figure 5-4 Sentiment classification model training flow chart

由图 5-4 可知，情感分类模型的训练是一个不断迭代的过程，模型训练过程为：

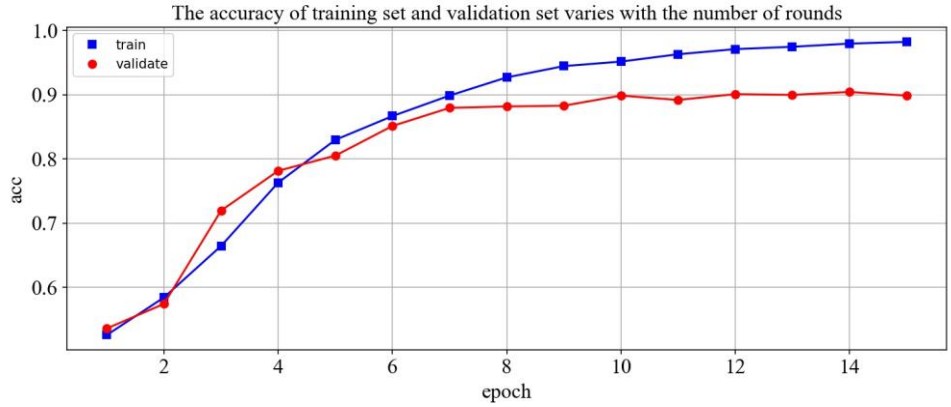
1) 将打标数据分为训练集和测试集，并输入到情感分类模型中；

2) 使用训练集训练模型，确定模型的训练轮数，分析模型在训练集和测试集上的准确率和误差的变化趋势并调整模型参数；

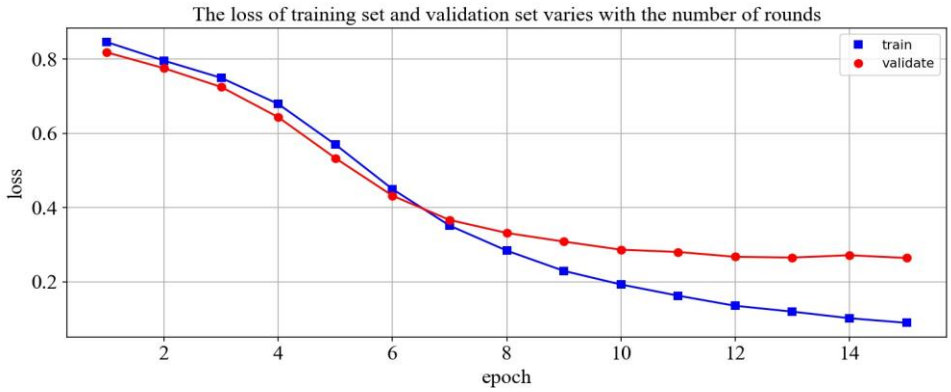
3) 不断调整模型参数，在不发生过拟合的情况下找到准确率最高、误差最小的模型，并将模型保存。

本文对每个主题在 2018 年 1 月的所有微博数据人工标注情感极性，标注方法如下：人工阅读待标注的每条微博数据，根据微博数据的语义判断该句子所表达的情感极性，将正面情感的微博数据标 1，负面情感的微博数据标 0，将人工无法判断极性和重复多次的微博数据舍弃，最终从 2018 年 1 月的 17494 条微博数据中标注出 8883 条数据，其中正面情感极性的微博数据有 4636 条，负面情感极性的微博数据有 4247 条。并按照 8:1:1 的比例随机选择其中的 7106 条微博数据作为训练集，888 条微博数据作为验证集，889 条微博数据作为测试集。

通过训练模型并不断调整参数，使模型的参数达到最优。在最优参数下，模型在训练集和验证集上的准确率和误差随着训练轮数的变化如图 5-5 所示：



a) 模型的准确率变化曲线
a) Accuracy rate curve of the model



b) 模型的误差变化曲线
b) Loss curve of the model

图 5-5 准确率和误差随着训练轮数的变化过程
Figure 5-5 Accuracy rate and loss as the number of training rounds changes

图 5-5 中，a)图是情感分类模型在训练集和验证集上的准确率随着训练轮数的

变化曲线，其中横坐标和纵坐标分别表示训练轮数和准确率；b)图是情感分类模型在训练集和验证集上的误差随着训练轮数的变化曲线，其中横坐标和纵坐标分别表示训练轮数和误差。

从图 5-5 中可以看出，在模型训练的前 13 轮，模型在训练集和验证集上的准确率随着训练轮数的增加不断增大，误差随着训练轮数的增加不断减小；从第 13 轮开始，虽然模型在训练集上的准确率不断增加，误差不断减小，但在测试集上的准确率和误差基本保持不变。因此模型在第 13 轮训练完毕，本文保存第 13 轮训练结束后的模型作为最终的模型。

使用保存的模型对测试集中的数据进行情感极性预测，达到了 90.45%的准确率。模型在测试集的部分数据上的预测结果如图 5-6 所示：

预测	真实	概率	原始句子
0	0	0.0	这么无懈可击的论点，竟然没有赢。我只能说现场100位都真奇葩
1	1	1.0	转发微博
0	0	0.02	呵呵
1	1	0.55	自改革开放以来，中国像是蛰伏在亚洲的一条巨龙，抬起了头，我为祖国有着如今的发展感到骄傲，这都得益于政府正确的发展政策和国人中精英人才的贡献。作为当代大学生，我们也应该努力提升自己，为祖国的未来做一番贡献！
1	1	0.96	这是真正的德艺双馨！
1	1	0.98	好
1	1	0.71	浪花与水同一物　老师所乐何　乐志于学 好于学也　凡志于学者皆有孤往精神

图 5-6 模型在测试集上的预测结果

Figure 5-6 The prediction result of the model on the test set

图 5-6 中，“原始句子”表示情感分类模型对这些句子的情感极性进行预测，“概率”表示模型对句子的预测值，该值大于 0.5 表示“原始句子”的情感极性为正面，小于 0.5 表示“原始句子”的情感极性为负面，“真实”表示句子的人工打标的标签，“预测”表示模型对句子的预测标签。

本文同时设计了三种基于传统机器学习算法的情感分类模型与基于双向 LSTM 算法的情感分类模型进行对比实验。由于传统机器学习模型需要对训练数据人工选择特征，综合考虑微博数据的特点和常用的特征选择方法的适用场景，本文选择卡方检验对机器学习情感分类模型的训练数据进行特征选择。这四种模型在测试集上的准确率如表 5-2 所示：

表 5-2 不同模型的实验结果对比

Table 5-2 Comparison of experimental results of different models

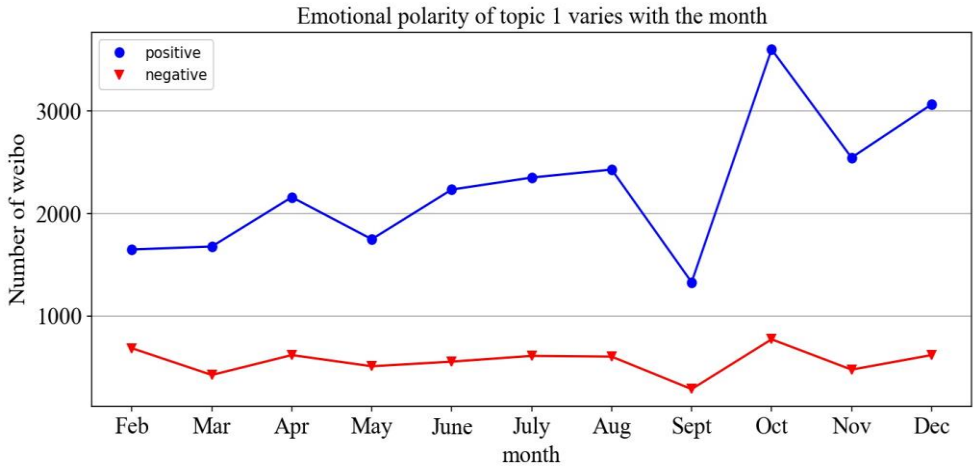
模型	测试集准确率
支持向量机模型	65.86%
逻辑回归模型	55.62%
朴素贝叶斯模型	55.94%
LSTM 模型	**90.45%**

从表 5-2 中可以看出，本文设计的三种基于传统的机器学习算法训练的情感分类模型在测试集上的准确率远远低于基于双向 LSTM 算法的情感分类模型。原因在于：微博数据的语言表达比较口语化，内容较复杂，传统的机器学习算法只能学习到数据中的线性特征，而双向 LSTM 算法可以学习到数据中的非线性特征，对复杂数据具有更强的刻画能力。

因此本文最终选择基于双向 LSTM 算法的情感分类模型来预测所有未标注数据的情感极性并自动标注。

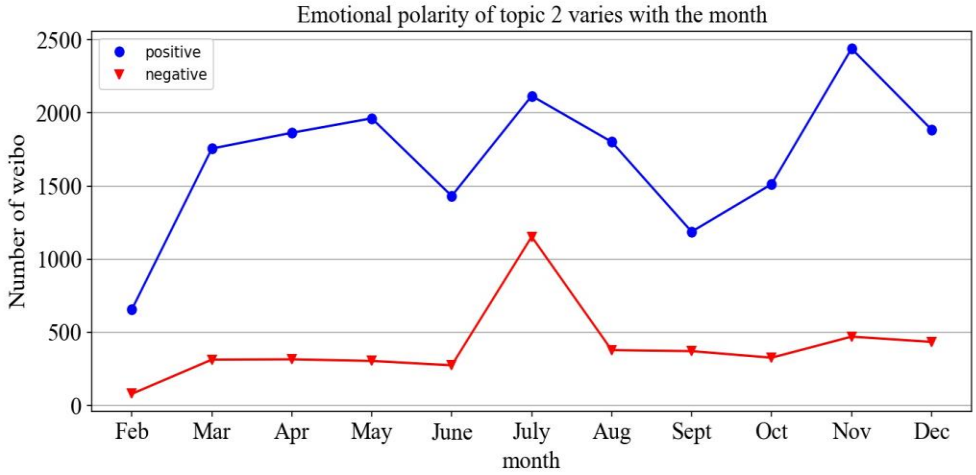
(2) 情感演进趋势分析

情感演进趋势分析统计每个主题在不同时间段的不同情感极性的微博数量，绘制不同情感极性的微博数量随时间的变化曲线。本文以月为单位统计每个主题在不同月份的正负面情感的微博数量，绘制情感演进趋势图，结果如图 5-7 所示：



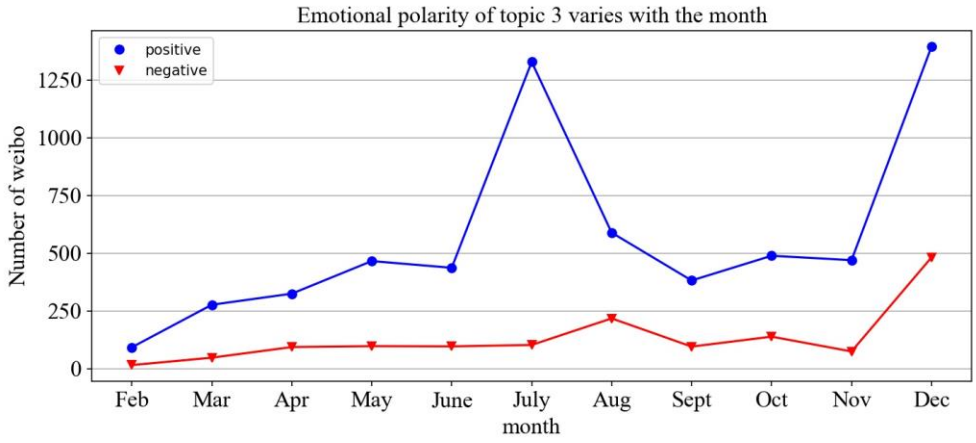
a) 主题 1 的情感变化曲线

a) Emotional curve of topic one



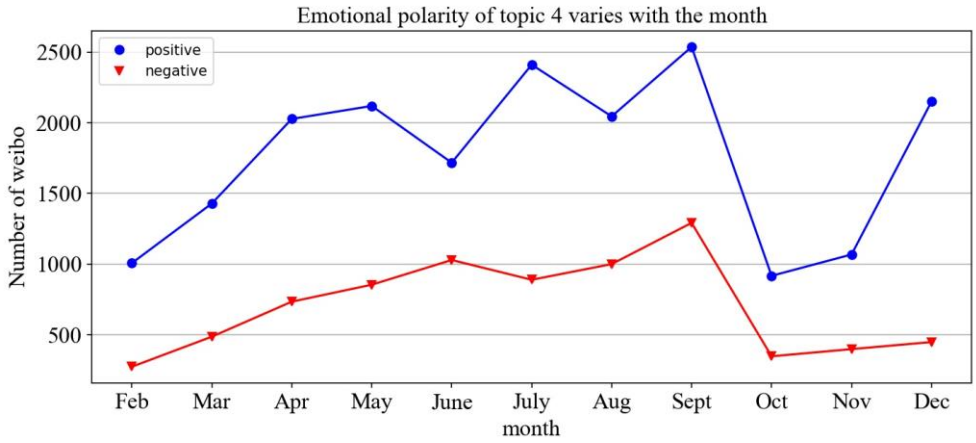
b) 主题 2 的情感变化曲线

b) Emotional curve of topic two



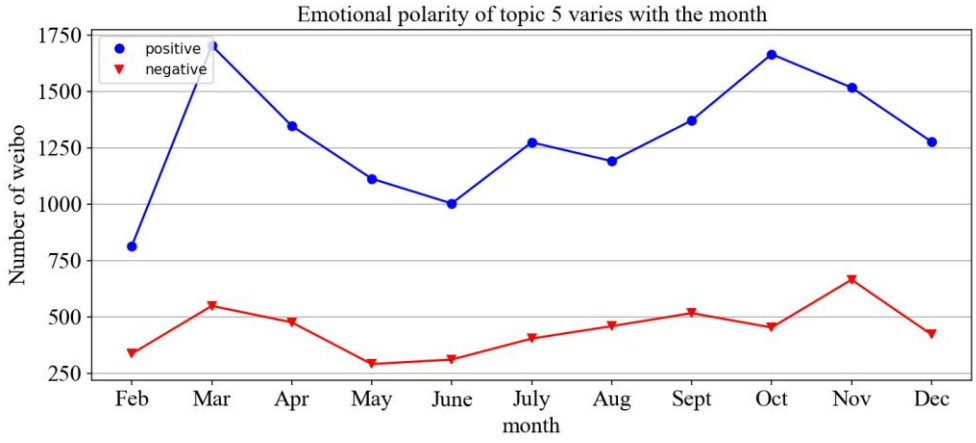
c) 主题 3 的情感变化曲线

c) Emotional curve of topic three



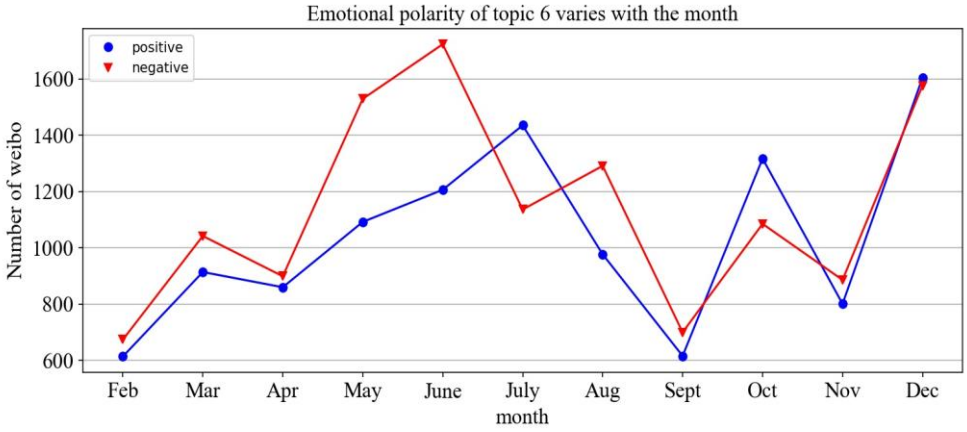
d) 主题 4 的情感变化曲线

d) Emotional curve of topic four

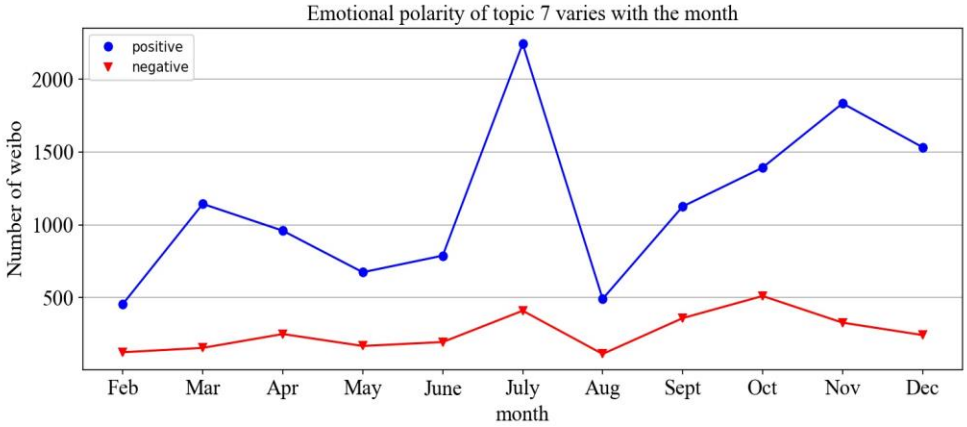


e) 主题 5 的情感变化曲线

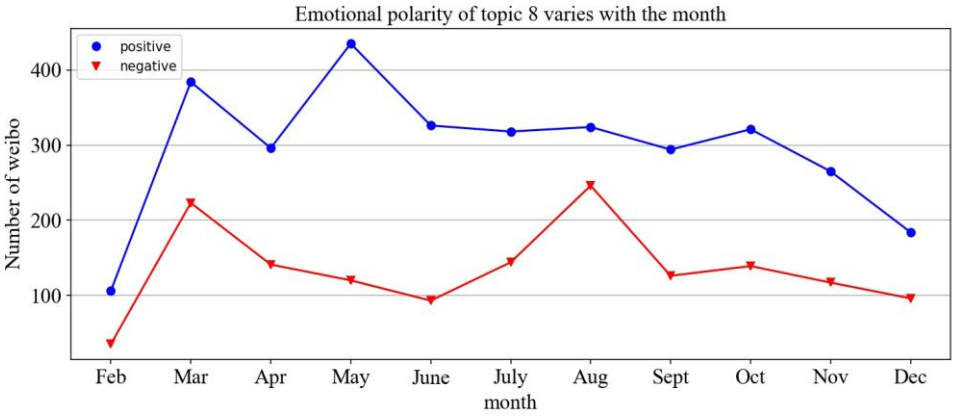
e) Emotional curve of topic five



f) 主题 6 的情感变化曲线
f) Emotional curve of topic six



g) 主题 7 的情感变化曲线
g) Emotional curve of topic seven



h) 主题 8 的情感变化曲线
h) Emotional curve of topic eight

图 5-7 民众对每个主题的情感变化曲线
Figure 5-7 People's emotional curve for each topic

图 5-7 表示每个主题对应的每个月的不同情感极性的微博数量随月份的变化曲线，蓝线表示正面情感的微博数量随月份的变化趋势，红线表示负面情感的微博数量随月份的变化趋势，其中横坐标表示月份，纵坐标表示微博数量。从图 5-7 中可以得出以下几个结论：

1) 总体上看，除主题 6 外，其余主题在每个月的正面情感的微博数均多于负面情感的微博数，说明人们对每个主题的提案所提到的内容是非常支持的，每个主题的提案是符合民意的；

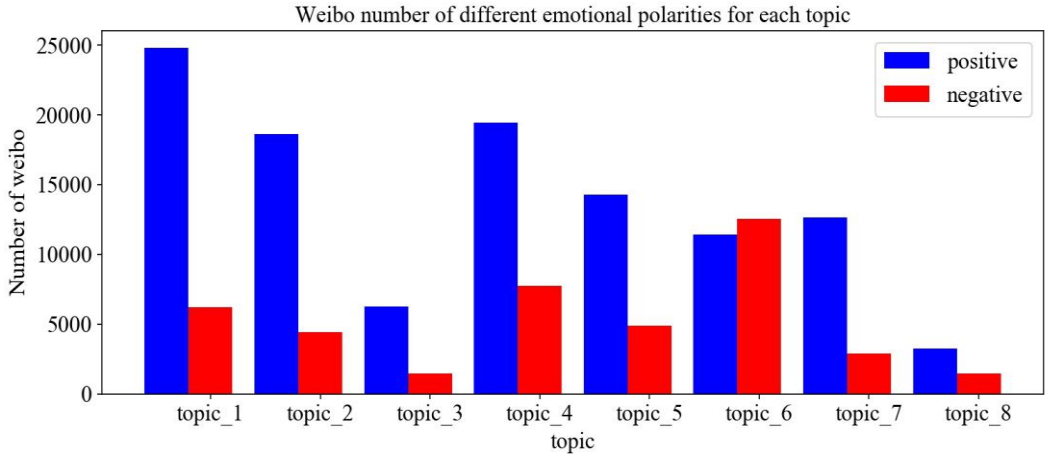
2) 主题 6 的关键词是[“共享单车”，“老年代步车”，“物业管理”]，该主题每个月正负面情感的微博数基本相同，说明人们对该主题喜忧参半，同时也反映了“共享单车”和“老年代步车”在使用和管理等方面可能存在的问题；

3) 从情感演进趋势上看，每个主题的正负面情感演进趋势基本一致，正负面情感的微博数量随着时间的变化几乎同时增加或减少，说明在 2018 年，虽然人们对每个主题的关注度不断变化，但相同时间段内正负面情感的微博数占该时间段微博情感总数的比例基本不变；

4) 虽然每个主题正负面情感的微博数量在政协提案提出后不断起伏变化，但总体上正面情感的微博数量随时间呈现上升趋势，负面情感的微博数量随时间呈现上升趋势或基本保持不变，说明人们不仅支持每个主题的政协提案提出的内容，而且支持率呈现上升趋势。

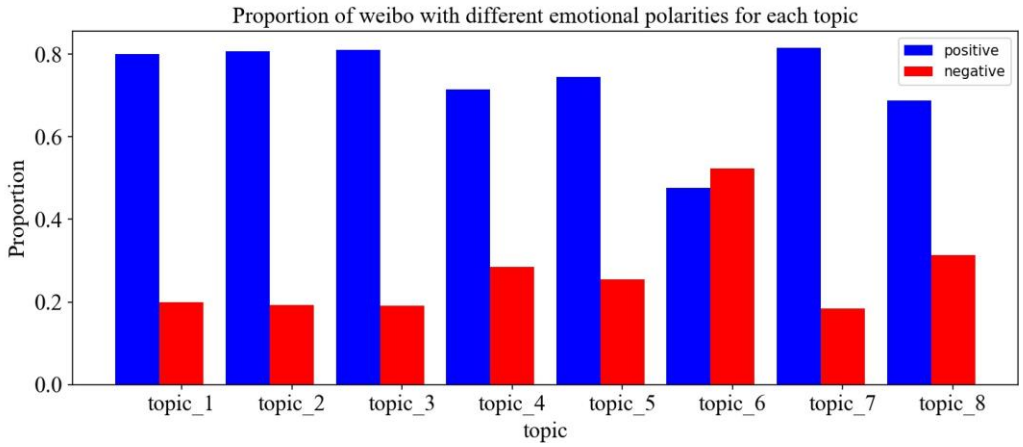
(3) 情感倾向总体分析

情感倾向总体分析指统计每个主题在 2018 年 2 月-2018 年 12 月的不同情感极性的微博数量，分析人们对每个主题的整体情感倾向。本文绘制了每个主题不同情感极性的微博数量柱状图和每个主题不同情感极性的微博数量占该主题微博总数的比例柱状图。结果如图 5-8 所示：



a) 每个主题不同情感极性的微博数量

a) Number of Weibos with different emotional polarities of each topic



b) 每个主题不同情感比例的微博数量

b) Number of Weibos with different emotional ratios of each topic

图 5-8 民众对每个主题的整体情感倾向

Figure 5-8 The general sentimental tendency of the people on each topic

图 5-8 中，a)图表示每个主题正负面情感极性的微博数量，其中蓝色柱状图表示每个主题正面情感极性的微博数量，红色柱状图表示每个主题负面情感极性的微博数量；b)图表示每个主题正负面情感极性的微博数量占该主题微博总数的比例，其中蓝色柱状图表示正面情感极性的微博数量占微博总数的比例，红色柱状图表示负面情感极性的微博数量占微博总数的比例。

从图 5-8 中可以得出以下几个结论：

1) 主题 6 负面情感的微博数量略多于正面情感的微博数量，其余主题正面情感的微博数均远远多于负面情感的微博数，说明人们对除主题 6 外的其余主题中的提案非常认可，对主题 6 中的提案仁者见仁、智者见智；

2) 虽然主题 1、主题 2、主题 3、主题 7 正负面情感的微博数量各不相同，但这四个主题正面情感的微博数约占每个主题微博总数的 80%，负面情感的微博数约占每个主题微博总数的 20%，正面情感的微博数是负面情感微博数的四倍，因此可以断定人们对这四个主题中的提案呈现一边倒的态度；主题 4、主题 5、主题 8 正面情感的微博数约是负面情感微博数的 2-3 倍，说明人们对这三个主题中的提案也非常认可；主题 6 正负面情感的微博数接近 1:1，说明人们对这个主题中的提案看法各不相同。

5.5 本章小结

本章首先介绍了本文的实验环境，然后对本文从每个主题中提取出的两组关

键词进行了有效性分析，由于人工标注关键词具有很大的主观性，因此本文设计实验从侧面对两组关键词的有效性进行了对比分析，接着介绍了从微博中采集每个关键词的微博舆情数据的方法，并展示了每个主题中每个关键词的微博数量，最后分析了人们对每个主题的关注度和情感，具体对每个主题从关注度演进趋势、关注度大小、情感演进趋势、情感总体倾向等四个方面进行了详细地分析。

6 结论

6.1 本文工作总结

全国政协提案是我国政治制度的重要机制之一，每年全国各级政协委员都要提出提案，然而形成提案需要花费大量的时间和精力。在互联网络时代，对政协提案进行热点主题发现和相关舆情分析能为政协提案提供辅助参考信息，为政协委员提供信息技术支持，从而节约政协委员形成新提案的时间和精力。本文以 2018 年北京市的政协提案为研究对象，分析提案所引起的社会舆情及变化趋势。主要进行了如下研究：

(1) 对政协提案划分主题并提取关键词。本文开发爬虫程序从北京市政协网站采集了 2018 年的全部政协提案数据，共采集提案 798 件，并保存为了结构化格式；探索了一种提案向量化表示方法，并使用 K-means 聚类算法将 2018 年北京市的 798 件政协提案聚成了 8 类，每一类代表一种主题；设计了两种关键词提取算法从每个主题中分别提取出了三个关键词，分别简称“长词”和“短词”，并设计对比实验分析了两组关键词的有效性，结果表明“长词”更能反映主题内容。

(2) 情感分类模型的设计、训练、数据标签预测。通过分析微博网页结构开发了微博爬虫程序，并解决了在爬取微博舆情数据的过程中遇到的问题，将从 8 个主题中提取出的 24 个“长词”采集的微博数据保存到了 Excel 表格中；基于双向 LSTM 算法设计了情感分类模型，将所有“长词”在 2018 年 1 月的微博数据进行人工打标，并按 8:1:1 的比例将其划分为了训练集、验证集和测试集，其中训练集中的数据用于训练模型，验证集中的数据用于辅助模型训练，通过观察模型在验证集上的准确率和误差来调节模型参数，测试集中的数据用于模型的准确率测试，最终在测试集上达到了 90.45%的准确率，远远高于基于传统机器学习算法的情感分类模型在该数据集上的测试准确率。

(3) 对政协提案的相关舆情进行统计并可视化。在上述工作的基础上，对获取的微博舆情数据进行了统计，主要分析了人们对每个主题的关注度演进趋势、关注度大小、情感演进趋势、情感总体倾向等四个方面。得出以下结论：在政协提案提出后，民众对每个主题的关注度总体呈现上升趋势；政协委员关注的热点主题也是民众最关注的主题；民众对几乎每个主题的正面情感都远远大于负面情感，且正面情感的增长速度大于负面情感。

6.2 未来工作展望

虽然本文的研究工作取得了一定的成果，但仍有一些不足之处，未来可以从以下几个方面对本文的研究工作进行改进：

(1) 由于政协提案向量的表示是否合理会影响主题划分效果，从而影响提取的关键词的质量，最终影响舆情分析效果，因此可以结合提案结构探索更合理的向量化方法。

(2) 本文从每个主题中提取的关键词虽然在主题上具有一致性，但使用这些关键词采集的微博数据仍会存在与主题不相关的数据，后期可以尝试不同的关键词提取方法。

(3) 本文只对比了基于双向 LSTM 算法的情感分类模型和基于传统机器学习算法的情感分类模型，未来可以尝试使用更多的深度学习算法设计情感分类模型，也可以将几种算法进行组合设计情感分类模型。

(4) 本文对政协提案的实施效果进行舆情分析只从关注度分析和情感分析两个方面展开，未来可以尝试从更多的方面进行分析。

参考文献

[1] 中国人民政治协商会议北京市委员会．http://tian.bjzx.gov.cn/goShowList.

[2] 王君扬．政协提案在基层政府决策中的作用分析[D]．华侨大学，2015.

[3] 张玉权． 谈谈提案的采纳与落实问题——关于提案办理评估体系的思考与探索[J]．湖北省社会主义学院学报，2007(6):36-37.

[4] 王法顺．政协提案信息化管理系统的设计与实现[D]．山东大学，2008.

[5] 赵琦， 张智雄， 孙坦， et al． 主题发现技术方法研究[J]． 情报理论与实践，2009, 32(4):104-108.

[6] Cheung D W , Kao B , Lee J . Discovering user access patterns on the World Wide Web[J]. Knowledge-Based Systems, 1998, 10(7):463-470.

[7] Perkowitz M , Etzioni O . Towards adaptive Web sites: Conceptual framework and case study[J]. Artificial Intelligence, 2000, 118(1-2):245-275.

[8] Mehrotra R , Sanner S , Buntine W , et al. Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling[C]// The 36th Annual ACM SIGIR Conference. ACM, 2013.

[9] 王李冬， 魏宝刚， 袁杰． 基于概率主题模型的文档聚类[J]． 电子学报， 2012, 40(11):2346-2350.

[10] 郭建永， 蔡勇， 甄艳霞． 基于文本聚类技术的主题发现[J]． 计算机工程与设计， 2008, 29(6):1426-1428.

[11] Luhn H P . A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. Ibm Journal of Research and Development, 1957, 1(4):309-317.

[12] Hulth A . Improved automatic keyword extraction given more linguistic knowledge[C]// Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2003.

[13] Fortuna, Blaž, Mladenič, Dunja, Grobelnik M . Semi-automatic construction of topic ontologies[M]// Semantics, Web and Mining. Springer Berlin Heidelberg, 2006.

[14] Xu S , Kong F . Toward better keywords extraction[C]// International Conference on Asian Language Processing. IEEE, 2015.

[15] 马力，焦李成，白琳，周雅夫，董洛兵．基于小世界模型的复合关键词提取方法研究[J]．中文信息学报，2009,23(03):121-128.

[16] 陈忆群，周如旗，朱蔚恒，李梦婷，印鉴．挖掘专利知识实现关键词自动抽取[J]．计算机研究与发展，2016,53(08):1740-1752.

[17] Subasic P, Huettner A. Affect analysis of text using fuzzy semantic typing[J]. Fuzzy Systems IEEE Transactions on, 2001, 9(4):483-496.

[18] Shanahan J G, Yan Q, Wiebe J. Computing Attitude and Affect in Text: Theory and Applications[M]. Springer Netherlands, 2006.

[19] Esuli A, Sebastiani F. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining[C]// 2006:417--422.

[20] 娄德成，姚天昉．汉语句子语义极性分析和观点抽取方法的研究[J]．计算机应用，2006, 26(11):2622-2625.
[21] Pang T B, Pang B, Lee L. Thumbs up? Sentiment Classification using Machine Learning[J]. Empirical Methods in Natural Language Processing, 2002:79-86.
[22] Bo P, Lee L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales[J]. 2005:115-124.
[23] Whitelaw C, Garg N, Argamon S. Using appraisal groups for sentiment analysis[J]. 2005, 9(1):625-631.
[24] 李思，张浩，徐蔚然, 等．基于合并模型的中文文本情感分析[C]// 全国信息检索学术会议．2009.
[25] 徐军，丁宇新，王晓龙．使用机器学习方法进行新闻的情感自动分类[J]．中文信息学报，2007, 21(6):95-100.
[26] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]// Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John Mcintyre Conference Centre, Edinburgh, Uk, A Meeting of Sigdat, A Special Interest Group of the ACL. DBLP, 2011:151-161.
[27] Tai K S, Socher R, Manning C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[J]. Computer Science, 2015, 5(1):: 36.
[28] Brueckner R, Schulter B. Social signal classification using deep blstm recurrent neural networks[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2014:4823-4827.
[29] 何炎祥，孙松涛，牛菲菲, 等．用于微博情感分析的一种情感语义增强的深度学习模型[J]．计算机学报，2017, 40(4):773-790.
[30] 梁军，柴玉梅，原慧斌, 等．基于深度学习的微博情感分析[J]．中文信息学报，2014, 28(5):155-161.
[31] Su K Y. A Character-Based Joint Model for Chinese Word Segmentation.[C]// International Conference on Computational Linguistics. Association for Computational Linguistics, 2010:1173-1181.
[32] Peng F , Feng F , Mccallum A . Chinese segmentation and new word detection using conditional random fields[J]. Proceedings of Coling, 2004:562--568.
[33] BOSON 中文语义开放平台．https://bosonnlp.com.
[34] 哈工大语言云技术平台．http://www.ltp-cloud.com.
[35] 结巴分词工具开源社区．https://github.com/fxsjy/jieba.
[36] 张健，杨淑芳．自然语言处理发展综述[J]．新教育时代电子杂志：教师版，2016(48).
[37] Moon T , Erk K , Baldridge J . Crouching Dirichlet, hidden Markov model: unsupervised POS tagging with context local tag generation[C]// Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010.
[38] Sun X , Ren F , Huang D , et al. Dual-chain Unequal-state CRF for Chinese new word detection and POS tagging[C]// International Conference on Natural Language Processing & Knowledge Engineering. IEEE, 2009.
[39] McCallum, Andrew, Dayne Freitag, and Fernando CN Pereira. "Maximum Entropy Markov Models for Information Extraction and Segmentation." Icml. Vol. 17. 2000.

[40] Wilbur W J , Sirotkin K . The automatic identification of stop words[J]. Journal of Information Science, 1992, 18(1):45-55.

[41] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[42] 周志华. 机器学习[M]. 清华大学出版社，2016.

[43] Krizhenvshky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional networks[C]//Proceedings of the Conference Neural Information Processing Systems (NIPS). 1097-1105.

[44] 李航. 统计学习方法[M]. 清华大学出版社,2012.

[45] 官琴， 邓三鸿， 王昊. 中文文本聚类常用停用词表对比研究[J]. 数据分析与知识发现, 2006, 1(3).

[46] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Machine Learning Research Archive, 2003, 3:993-1022.

[47] Hofmann T . Probabilistic Latent Semantic Analysis[J]. 1999.

[48] Deerwester S , Dumais S T , Furnas G W , et al. Indexing by latent semantic analysis[J]. Journal of the Association for Information Science & Technology, 2010, 41(6):391-407.

[49] Le Q V , Mikolov T . Distributed Representations of Sentences and Documents[J]. 2014.

[50] 杨春明， 韩永国. 快速的领域文档关键词自动提取算法[J]. 计算机工程与设计， 2011, 32(6):2142-2145.

[51] Hulth A. Combining machine learning and natural language processing for automatic keyword extraction[M]. Department of Computer and Systems Sciences [Institutionen för Data-och systemvetenskap], Univ., 2004.

作者简历及攻读硕士/博士学位期间取得的研究成果

一、作者简历

刘一健，男，汉族，1995 年出生，祖籍山西。2013 年 9 月-2017 年 7 月就读于中北大学，获得工学学士学位，2017 年 9 月-2019 年 6 月就读于北京交通大学，获得专业硕士学位。

二、发表论文

[1]

[2]

[3]

.

.

.

三、参与科研项目

[1]参与实验室和北京市计算中心合作的科研项目——《基于大数据的提案系统及其实施效果相关舆情演进分析》

[2]

[3]

.

.

.

四、专利

[1]

[2]

[3]

.

.

独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作和取得的研究成果，除了文中特别加以标注和致谢之处外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得北京交通大学或其他教育机构的学位或证书而使用过的材料。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：刘一健 签字日期：2019 年 5 月 27 日

学位论文数据集

表 1.1： 数据集页

<table>
<tr><td>关键词*</td><td>密级*</td><td>中图分类号</td><td>UDC</td><td>论文资助</td></tr>
<tr><td>主题划分、关键词提取、爬虫、情感分类、舆情分析</td><td>公开</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">学位授予单位名称*</td><td>学位授予单位代码*</td><td>学位类别*</td><td>学位级别*</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>专业硕士</td><td>硕士</td></tr>
<tr><td colspan="2">论文题名*</td><td colspan="2">并列题名</td><td>论文语种*</td></tr>
<tr><td colspan="2">基于机器学习的政协提案和相关舆情的分析</td><td colspan="2"></td><td>中文</td></tr>
<tr><td>作者姓名*</td><td colspan="2">刘一健</td><td>学号*</td><td>17125039</td></tr>
<tr><td colspan="2">培养单位名称*</td><td>培养单位代码*</td><td>培养单位地址</td><td>邮编</td></tr>
<tr><td colspan="2">北京交通大学</td><td>10004</td><td>北京市海淀区西直门外上园村 3 号</td><td>100044</td></tr>
<tr><td colspan="2">工程领域*</td><td>研究方向*</td><td>学制*</td><td>学位授予年*</td></tr>
<tr><td colspan="2">电子与通信工程</td><td>信息网络</td><td>2 年</td><td>2019 年</td></tr>
<tr><td>论文提交日期*</td><td colspan="4">2019 年 5 月 27 日</td></tr>
<tr><td>导师姓名*</td><td colspan="2">赵永祥</td><td>职称*</td><td>副教授</td></tr>
<tr><td>评阅人</td><td colspan="2">答辩委员会主席*</td><td colspan="2">答辩委员会成员</td></tr>
<tr><td></td><td colspan="2">孙强</td><td colspan="2">郭宇春、李纯喜、张立军、郑宏云</td></tr>
<tr><td colspan="5">电子版论文提交格式 文本（ ） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ）
推荐格式：application/msword；application/pdf</td></tr>
<tr><td colspan="2">电子版论文出版（发布）者</td><td colspan="2">电子版论文出版（发布）地</td><td>权限声明</td></tr>
<tr><td colspan="2"></td><td colspan="2"></td><td></td></tr>
<tr><td>论文总页数*</td><td colspan="4">63 页</td></tr>
<tr><td colspan="5">共 33 项，其中带*为必填数据，为 21 项。</td></tr>
</table>