摘要

网络为公众提供了更开放平等的交流平台的同时，使公众能随时获取和分享一手资讯。但网络是一把双刃剑，不仅会引发大规模的舆情危机，也能成为社会情绪的减压阀。在互联网时代，网络逐渐成为信息传播的主要媒介之一，信息通过网络传播的速度越来越快，公众通过网络参与话题讨论的频率也越来越高。其中，突发公共事件在网络传播的过程中更容易演变成舆情危机，如果媒体人及政府及时发现问题，澄清谣言并疏导公众负面情绪，就能将其遏制在萌芽状态。

鉴于此，为帮助政府及媒体人更高效地开展网络舆情监管工作，本文搭建了一种混合模型，该模型在挖掘评论文本主题的同时，探究网民的情感走势，以提升舆情分析的全面性和准确度。在主题挖掘方面，该模型借助层次狄利克雷混合过程模型（HDP-vMF）细粒度提取微博评论中的热门主题。HDP-vMF 将每条评论数据视为独立的小组，基于组数据聚类法进行细粒度主题提取，并将 Gaussian 分布替换为 von-Mises Fisher 分布使得所有的簇被均匀映射到一个单位超球面上，而不是鼓励所有的簇都向原点聚集，从而减小聚类时产生的误差。在情感演化分析方面，利用深度迁移学习模型（BERT）对微博评论数据进行细粒度情感分类建模。由注意力网络组成的 BERT，在利用并行运算显著地提升运行速度的同时，多头注意力机制也增强了模型语义分析理解的能力，为评论情感分析研究提供了更加匹配的理论算法。

在实证分析方面，本文以“河南暴雨”事件为例，抓取 2021 年 7 月 19 日至 7 月 31 日的相关话题微博评论数据，共计 120791 条。初步清洗数据后，先利用 jieba 工具对数据进行分词处理，并使用 word2vec 与 BERT 将分词后的数据转换为词向量；然后，通过加权平均词向量获得句子向量，使用 BERT 实现评论数据的情感分类，并基于此绘制情感演化图；最后，借助 HDP-vMF 模型提取评论数据的主题词，得到微博用户关注的热点话题。实证分析结果表明：（1）情感演化方面：在舆情爆发期，微博用户的负面情绪居多，但在整个周期内，总体情绪倾向为正向。（2）热议话题方面：“河南暴雨情况”为微博用户讨论的核心话题。随着事件的进一步发展，衍生了多个热点话题，其中，“地铁四号线被淹”“明星及企业网络捐款”“K599 列车被困”“求救文档”为热度排名前四的话题，经与微博热搜话题验证，HDP-vMF 能有效地挖掘真实主题。（3）BERT 性能方面：通过将 BERT 与 6 种学术界常用的情感分类算法进行性能验证，BERT 的 ACC 及 F1 值均位列第一，验证了 BERT 在情感分类方面的优越性能。

关键词：BERT；HDP-vMF；主题挖掘；情感演化；河南暴雨

Abstract

The Internet provides a more open and equal communication platform for the public, while enabling the public to access and share first-hand information at any time. However, the Internet is a double-edged sword, which will not only trigger a large-scale public opinion crisis, but also become a relief valve for social emotions. In the Internet era, the network has gradually become one of the main media of information dissemination. The speed of information dissemination through the network is faster and faster, and the piblic participates in topic discussion through the network more and more frequently. Among them, in the process of network communication, public emergencies are more likely to evolve into public opinion crisis. If the media and the government find problems in time, clarify rumors and dredge the public's negative emotions, they can be contained in the bud.

In view of this, in order to help the government and media managers carry out the supervision of online public opinion more efficiently, this paper establishes a hybrid model. While mining the theme of comment text, the model explores the emotional trend of Internet users in the whole stage of public opinion, so as to improve the comprehensiveness and accuracy of public opinion analysis. In terms of topic mining, the model uses the hierarchical Dirichlet mixed process model (HDP VMF) to fine-grained extract the hot topics of microblog comments. HDP VMF regards each comment data as an independent group, extracts fine-grained topics based on group data clustering method, and replaces Gaussian distribution with von Mises Fisher distribution, so that all clusters are evenly mapped to a unit hypersphere, rather than encouraging cluster centers to converge to the original point, so as to reduce the error in the clustering process. In the aspect of emotion evolution analysis, the deep transfer learning model (BERT) is used to model the fine-grained emotion classification of microblog comment data. BERT, which is composed of attention network, significantly improves the running speed by using parallel operation. At the same time, the multi head attention mechanism also enhances the ability of semantic analysis and understanding of the model, and provides a more matching theoretical algorithm for the research of critical emotion analysis.

In terms of empirical analysis, taking the "rainstorm in Henan" event as an example, this paper captures the data of microblog comments on relevant topics from

July 19 to July 31, 2021, with a total of 120791. After the preliminary cleaning of the data, the jieba tool is used to segment the data, and the word2vec combined with BERT is used to convert the segmented data into word vectors. Then, the sentence vector is obtained by weighted average word vector, and the sentiment classification of comment data is realized by BERT, and based on this, the sentiment evolution diagram is drawn. Finally, with the help of HDP-vMF model, the key words of the review data are extracted, and the hot topics concerned by users in Weibo are obtained. The empirical results show that: (1) Emotion evolution: In the period of public opinion outbreak, most of the negative emotions of Weibo users, but the overall emotional tendency in the whole cycle is positive. (2) Hot topics: "rainstorm in Henan" is the core topic discussed by microblog users during the period of public opinion. With the further development of the incident, a number of hot topics have emerged, among which "Metro Line 4 flooded", "star and enterprise network donations", "k599 train trapped" and "rescue documents" are the top four topics， HDP-vMF can effectively mine real topics through verification with Weibo hot search topics. (3) BERT performance: through the performance verification of BERT and six commonly used emotion classification algorithms in academic circles, BERT's ACC and F1 values rank first, which verifies BERT's superior performance in emotion classification.

Keywords：BERT；HDP-vMF；theme mining；emotion evolution；Henan heavy rain

目 录

第 1 章 绪论

1.1 研究背景及意义

与 2003 年非典疫情、2008 年汶川大地震发生的时代不同，随着互联网技术的蓬勃发展，以知乎、微信、微博、推特为代表的社交网络逐渐成为舆情传播的主要途径之一[1]。以 2021 年发生的重大自然灾害-河南特大暴雨为例，暴雨发生后，有关河南暴雨的微博话题热度数日居高不下，公众对河南暴雨相关的话题进行了上亿次讨论和转发，借此表达看法、抒发情感。这种以微博为代表的新兴社交软件的出现正在重塑网络舆情传播的结构，该类应用极具特色的“对话式”评论、连续转发分享功能不但可以调动更多用户参与各类事件、话题的讨论过程，也可能使某一话题上升到社会问题，进而引发大规模网络舆情事件。

近年来，突发公共事件频出，这类事件因为与公众的人身安全密切相关，而受到高度重视，话题热度持续时间久，难以在短期内平息[2]。在事件发展过程中，极易衍生出其他话题，网络舆论走向随时可能偏离正轨，引发公众负面情绪，造成不良社会影响，增加相关组织机构监管难度[3]。

近年来，机器学习和深度学习的发展，尤其是评论文本挖掘相关技术的日渐成熟，为突发公共事件的网络舆情监管以及预测提供了更加便捷可靠的途径[4]。在评论文本中，由于同一词汇所处语境不同，当多个词汇组合顺序发生改变时，文本传达出的语义也不尽相同。在机器学习和深度学习领域，文本作为模型输入的前提条件是要将词汇转换成可用于计算的词向量，而如何构建含有丰富语义信息的词向量则成为决定模型性能的关键。

目前，大多数词向量的构造往往使用词汇级别静态词向量 One-hot Vector，SVD 以及 word2vec，但上述方法一旦面临语境、语序等变化时，很难有效地将原始词汇之间的语义信息嵌入到词向量，动态处理性能不足[5]。另外，不同模型的架构也影响着情感分析与主题提取的相关性能。在情感分析方面，传统的情感分析方法仍存在一些问题[6]，例如，情感分类的性能往往受限于是否能从句字中挖掘更加完整的信息，如果能够考虑句子的全局信息、结合语境，则能促进模型较好地理解语义，从而实现更好的情感分类。现有的文本聚类和评论情感分析主要通过 K-means 算法[7]、TF-IDF 模型[8]、基于概率分布的 LDA 模型[9]、LSTM 模型[10]等实现。尽管上述学习方法相对较成熟，但依旧存在着明显的不足之处，例如：在利用 GaussianLDA 和 LDA-ARMA 模型进行聚类分析时，模型通常将所有词汇看作一个整体，此做法不但会丢失不同评论之间的关联性，而且无法细

粒度地提取不同类别的主题信息。同时，TF-IDF 模型虽然具有实现简单，易于理解等诸多优势，但它的算法精度不高的缺点也显而易见[11]。在实验过程中 TF-IDF 对语料库的依赖程度过高、向文本中频率较小的词语的倾斜特征明显。还有 LSTM 及其变种虽然能够很好地根据上下文信息更新参数，但是缓慢的推理速度无法满足生产环境下的延迟要求。在主题挖掘方面，传统的 LDA 主题模型在挖掘网络评论等短文本的主题方面存在一定局限性[12]，不少学者在主题模型基础上提出了改进措施，以此达到相应的实验目的，但是目前相关的主题模型并未考虑多条评论之间的语义关系。

为弥补上述不足，本文提出基于 von-MisesFisher 分布的层次狄利克雷混合过程模型（HDP-vMF）[13]和深度迁移学习模型(BERT)[14]。同时，以 2021 年河南暴雨舆情微博在线评论为例，展示本文提出的模型针对舆情不同时间维度热门话题的挖掘能力及公众情感动态变化过程的分析能力。该方法利用 BERT 强大的语义抓取功能和快速的逻辑推理，为评论的情感分析匹配更加精准的理论算法。通过 BERT 在特定领域微调后获得的含有语义信息的词向量结合 word2vec 的词向量构建更具多样性的输入，使用 HDP-vMF 模型的层次性质，从海量微博评论数据中并行提取更加细粒度的主题信息。基于此，本文旨在构建性能更优的混合模型，为媒体人与相关管理部门舆情监管工作提供更精准的热点话题与情感演化数据。

1.2 国内外研究综述

1.2.1 情感分类研究现状

情感分类是指结合实际情景，通过分析推理判断文本的感情色彩倾向，即评论者的态度是正向、负向或不带任何感情色彩的客观阐述，其属于情感分析领域的问题[15]。相比于传统的文本分类方法，情感分类主要的目的在于从评论数据或相关数据中挖掘到能支持某种观点的信息。因此，主要被用于探究文本中所呈现的主观态度。目前，情感分类研究主要有三种方法：情感词典分类、机器学习方法和深度学习方法[16]。

（1）情感词典分类方法

基于人工手动构建特定领域的词典是最基础的情感分类方法，符合人们对某事物的客观认知。目前，构建情感词典的方法为人工构建情感词典、自动构建情感词典[17]。充足的数据是人工构建情感词典的必要条件，因此需要抓取大量文本数据；然后，人工标注文本中每个词语的情感类别；最后，综合文本中所有词语的情感类别，即可判断该文本的情感倾向。大多数情感词典基本都只涵盖了常用

的词语，如果学者研究的是某些专业领域，则该领域的专业词汇可能需要另外补充。虽然人工构建词典便于扩充词条信息，但时间成本、人工成本较高，并且设计的范围有限，不适合跨领域研究[18]。自动构建情感词典是基于人工构建的方法，结合知识语料库或知识库为原有语料库添加特定领域词汇，进一步丰富语料，扩大机器情感分类的识别范围。为了使现有词典更加匹配自身研究领域，大多数学者都选择第二种词典构建方法。刘晓娟等[19]根据评论语料库自行构建信息公开舆情词典，成功弥补了 SnowNLP 自带语料库的滞后性与局限性。谭旭等[20]采用机器学习与情感词典结合的方法，探究香港“修例”风波中网民的情感演化过程，实证分析结果表明模型预测效果较好，平均误差低于 9.95%。Taboada 等[21]构造了一种带有语义方向注释的单词词典，并将其应用于极性分类任务，即为文本指定正面或负面标签的过程。Hu 等[22]提出了基于词频的词典构建方法来进行词性标注，核心思想在于通过从句子中抽取出观点词来进行情感分类。

（2）基于机器学习的情感分类方法

机器学习是利用现有的数据构建模型从而处理或预测未来的数据，所以机器学习是需要训练的[23]。在模型训练前期，先要完成文本特征处理工作，将重要的文本特征加入训练，以此提高模型评估指标。在模型训练过程中，采取有监督的训练模式，通过反复调整试验参数，提高模型预测性后准确预测实验对象。在模型训练结束后，便可用于新文本的情感极性预测。陈震等[24]利用贝叶斯模型（BN）分析网络舆情事件趋势，准确预测出网络舆情事件发展趋势，显著提高了网络舆情数据的准确性。孙松涛等[25]提出了一种表示方法，该方法利用多标签来区分不同特征，借助卷积神经网络（CNN）有效构建出具有一定情感语义差别的句子向量，多个标签分类器的分类性能均显著提高。P.S 和 Mahalakshmi[26]应用了朴素贝叶斯机器学习分类器，通过不同的情感类别对文本进行分类，基于文本开发了一个情感状态识别系统。为了从不同的情感维度进行更加细致地分析，Ghosh 等[27]应用决策树和朴素贝叶斯等多个机器学习分类器，利用集成学习的方式进一步提升了情感分类的准确率。

（3）基于深度器学习的情感分类方法

与机器学习不同，深度学习不需要人工设计规则训练模型学习，只需要输入数据，它就可以自己找到一些特征，然后进行分类学习。另外，深度学习能够在开发过程中收集大数据集对效能评估的系统性集成，成功解决了许多过去未解决的问题。贵向泉等[28]提出 TCN-BA 文本情感分析混合模型，该模型引入自注意力机制有效地优化了模型特征向量，改善了大多数神经网络无法充分利用上下文信息的缺陷，提高了情感分类准确度。马远等[29]在 SemEval2014 的两个数据集上进行了对比实验，分别获得了 81.33%和 74.22%的准确率，结果表明该双向注

意力机制可适用于更细粒度的文本情感分析。庄穆妮等[30]在 BERT 模型预训练任务的基础上叠加了深度预训练任务，并通过结合主题优化特征信息向量与 BERT 词向量来提高情感分类的精确度，试验数据表明：该模型使复杂文本情感分类的 AUC 值超过 99.6%。Qian 等人[31]将基于词汇的线索纳入基于 LSTM 模型的训练中，构建了基于深度网络的情感分类方法。其提出的方法依赖于一个新的损失函数，该函数考虑极性词或某些类型的词（例如 privative）与输入文本中相邻词之间的关系。Shin 等人[32]通过学习极性词的词汇表征来作为深度网络的辅助信息，并结合情感分类的词汇表征进一步优化了整体网络的分类效果。

1.2.2 主题模型研究现状

主题模型是文本挖掘的重要工具，一直以来备受诸多领域研究者关注。文本挖掘这一领域的大多数据集都以非结构化来呈现，很难从中直接获取有用信息，而主题模型的优势则是能够充分挖掘到语料里的基本表述信息，并且在主题聚合、特征选择等场景得到广泛应用[33]。目前，利用主题模型进行文本挖掘主要集中在事件监控与预测、技术方法和商业领域等方面。

在技术方法研究方面，陈伟等[34]利用 LDA 模型挖掘专利文献中潜在的技术主题，基于这些主题挖掘出了不同时期下技术的演变规律和分布特征，为专利计量提供便捷途径。关鹏等[35]以科学文献作为数据来源，使用 LDA 模型对关键词、摘要、关键词+摘要三种语料构成的语料库进行主题挖掘，基于挖掘出的相关主题，发现不同的语料对挖掘效果起到的影响不同。陈斌等[36]在 LDA 模型中引入热度值，构建了一种 HLDA 网络文本挖掘模型，其中热度值在模型中起到一个权重的作用，同时以更加直观的方式展示出高准确度的主题，并增强了模型的语义可解释性以及主题挖掘能力。Aytug Onan[37]提出了一种采用两个阶段的主题提取方案，第一步为文本的表征学习，即预先通过 word2vec、POS2vec、word-position2vec 和 LDA2vec 等方法得到的单词向量；之后将所有向量结合作为聚类方法的输入，第二步为主题聚类，其搭建的基于多种聚类算法的架构有效地实现了主题挖掘。Fan 等[38]提出了一个基于层次聚类的主题抽取方法，该方法基于分层贝叶斯非参数框架，允许语料库中不同的新闻故事共享主题。此外该方法基于分层 Pitman-Yor 过程混合模型，以逆 Beta-Liouville（IBL）分布作为其成分密度，该模型在文本数据建模方面表现出优于被广泛使用的高斯分布的性能。Andrzejewski 等[39]构建了一种新的 Dirichlet 森林先验知识，将特定领域知识整合到一个潜在的 Dirichlet 分配框架中，并通过折叠吉布斯抽样进行推断，实验表明该模型能够跟踪和概括用户指定的领域知识。阮光册[40]基于网络在线评论文本

短、信息量少的特征，提出一种将 HowNet 知识库与主题模型融合的方法挖掘文本信息，实验结果表明，该方法成功地提高了文本挖掘的准确性和实用性。

在商业领域研究方面，国显达等[41]在传统 LDA 主题模型的基础上提出 GaussianLDA 在线评论主题挖掘方法，挖掘天猫、豆瓣和京东平台中的用户评论，在实验过程中充分分析语义一致性问题，并基于此进行了相关调优，使生成的主题分布更紧密、语义更连贯。杨鑫等[42]使用情感词典与 SnowNLP 技术挖掘携程网站中的在线评论，以期探究游客真实诉求，帮助各景区负责人制定更加合理的经营策略。张艳丰等[43]为探寻网购者发布评论的普遍时间规律和关注重点，利用语义分析、情感分析和词频共现等方法挖掘在线评论文本和追加评论文本，从而为网购者提供更好的选购建议，为商家所售卖商品的后续迭代提供参考。李可、陈光平[44]使用深度学习 STV 模型学习在线评论潜在语意特征，并将深层语意特征挖掘模型与 SVD 模型结合，挖掘用户情感偏好。试验结果表明，该模型性能明显优于传统评论挖掘模型，能够向网购者提供更加精准的意向商品。

在事件监测方面，葛琳等[45]以主题模型为基础，建立了实时多维信息的在线信息内容安全事件分类模型（RMIA-LDA）。该模型不仅能够准确地完成信息内容安全事件的分类，还有效提高了相关任务的性能。曹树金和岳文玉[46]使用生命周期理论、TF-IDF 和潜在 LDA 模型的方法，有效挖掘了舆情传播周期中不同阶段的热点主题，旨在为舆情决策与分析提供科学依据。Mimno 等[47]定义了 Weblog 上的主题情绪分析问题，并提出主题−情绪混合（TSM）概率模型揭示博客集合中潜在的主题方面、特别查询结果中的子主题及其相关情绪。它还可以提供适用于任何特定主题的一般情绪模型。在不同的网络日志数据集上进行的实证实验表明，该方法可以有效地建模主题方面和情绪，并从网络日志集合中提取其动态。

1.2.3 研究现状评述

通过对情感分类、主题模型的国内外调研发现，现有的研究内容较为广泛，包括基于现有模型的改进、新模型的构造和实际应用研究，虽然还有一些问题仍待解决，但总体而言取得的成果十分显著。

早期虽然大多数国内外学者注意到挖掘在线评论的潜在价值，但相关研究主要集中于影评[48]、电商[49]等商业领域。近年来，由于社交软件成为主流的信息传播渠道，大规模网络舆情事件频发，探究网民情感趋势、挖掘关注话题变得尤为重要，因此，越来越多的学者开始研究舆情在线评论，改善了舆情评论研究不足的情况。另外，在舆情的不同发展阶段中，公众讨论的主题往往呈现多元化趋势，不同主题引发的情感变化也不尽相同，主题挖掘和情感演化分析相结合的分

析方法有助于政府和媒体人更加细致、准确、全面地开展舆情监管工作。现有评论文本挖掘研究往往只单一进行主题挖掘或静态情感分析，既忽略了舆情话题中主题和情感的关联性，又没有考虑舆情发展过程中公众的情感波动。

在主题挖掘方面，由于目前各个行业的数据体量迅猛增长，而在传统的 LDA 主题挖掘方法中，其浅层的模型结构无法有效地拟合大批量的数据，并且文本词汇的表征信息好坏决定了主题挖掘整体性能的上限，因此，现有主题挖掘方向的核心研究点主要通过深度神经网络与传统 LDA 相结合的方式训练建模。加入深度神经网络建模的原因在于神经网络层数具有可叠加性，充足的内部参数能有效地拟合大体量的数据分布。另外，神经网络还是一个非常有效的数据映射工具，能够将文本从高维空间中映射到一个稠密的低维空间，从而进一步提升主题挖掘的有效性。在情感分析方面，情感分析作为文本挖掘领域的基础任务之一，在整个链路中起着至关重要的作用，例如对电影评论数据进行情感挖掘后，电影制片人可以更准确地判断公众喜好，找准市场发展方向；对微博评论进行情感分析，媒体人可根据分析结果进行舆论引导，避免引发舆论危机。传统的分析方法主要为基于辞典与基于机器学习建模，基于辞典的方法需要人为地构造相关情感趋势的词汇，误差率较大；基于机器学习的方法无法有效地处理海量的数据。因此目前在情感分析方向的研究热点主要基于深度学习方法，尤其是基于 NLP 领域的预训练大模型，例如 BERT，Transformer 等，由于大模型能有效地抓取文本内的语义信息，从而可以更加细粒度的挖掘文本所表示的情感极性。此外，为了进一步增强深度学习的可解释性，有诸多研究方法结合深度网络与传统词典，通过嵌入辞典作为模型辅助信息进一步提升情感分析的性能。

综上所述，在情感演化分析方面，本文通过使用基于双向 Transformer 网络的 BERT 模型，利用多头注意力机制对句子进行全局分析解读，以此提高语句情感分类准确率。在主题挖掘方面，本文将狄利克雷过程（DP）扩展为层次狄利克雷过程（HDP），利用层次的性质来提取多条评论之间的语义关系，共享多个聚类主题。

1.3 研究内容和方法

1.3.1 研究内容

本文的技术路线图如图 1.1 所示：

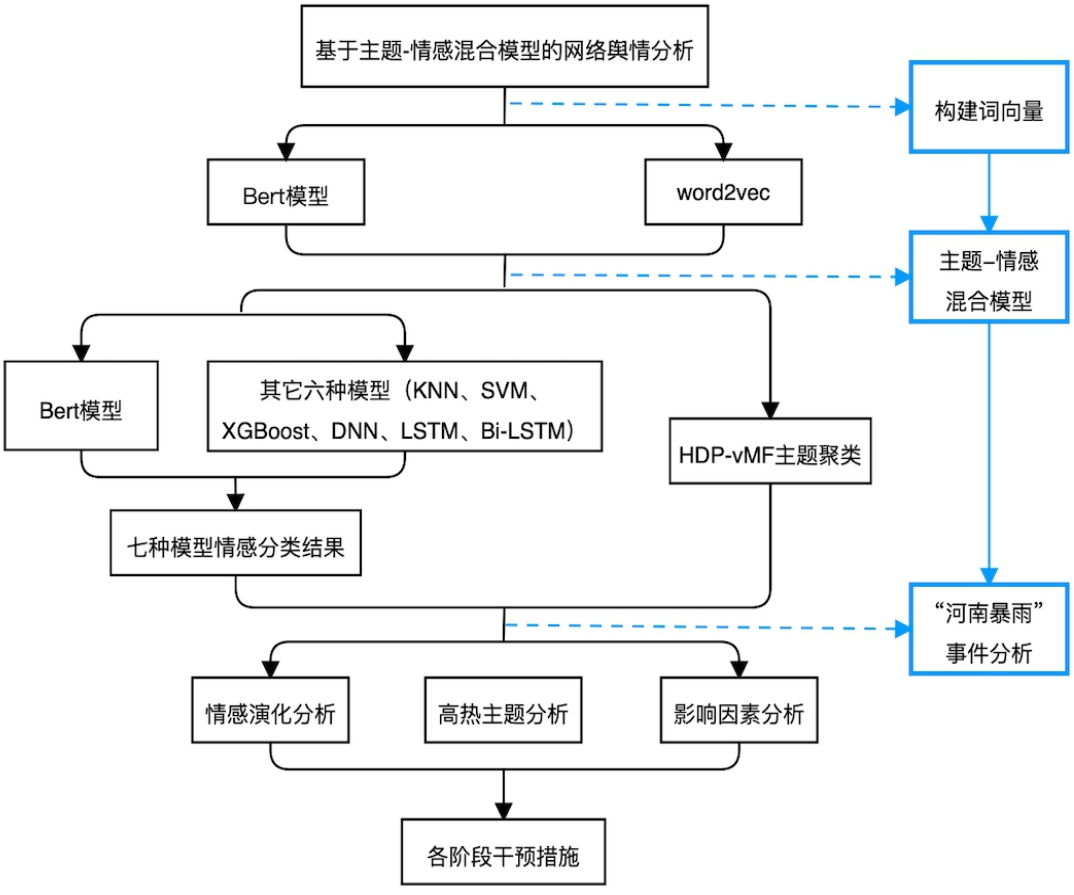

图 1.1 技术路线图

本文以新浪微博为载体，对突发公共安全事件的相关网络舆情进行情感趋势研究，并对微博评论文本进行分词、构建带有感情倾向的词向量、情感演化分析、热点话题聚类等。由于适合做主题挖掘与情感分析的模型众多，且这些模型的优缺点各不相同，因此，为了验证本文所采取的 BERT-HDP-vMF 混合模型效果更优，本文在确保其它变量均一致的情况下，选取了 6 种学术界常用的情感分类算法，进行对照试验。同时，为了验证该模型的准确度，本文选取 2021 年 7 月 19 日发生的“河南暴雨”事件进行实证分析。

本文主要研究内容归纳如下：

一是关于“河南暴雨”事件的微博评论数据获取与预处理。由于模型训练需要借助大量微博数据做支撑，因此本文基于 python 中提供的爬虫工具，以关键词的方式来对评论进行匹配爬取，利用正则表达式进行特殊字符匹配删除表情符号以及标点符号，并通过分词、去除停用词等操作完成原始数据的预处理工作。

二是使用预处理后的数据构建带有情感倾向的词向量。为获取情感值更加准确的词向量，本文将全部数据分成两组，第一组数据用于 word2vec 构建词向量、第二组数据用于微调后的 BERT 模型构建词向量。

三是根据“河南暴雨”事件的舆情发展阶段，进行主题挖掘及情感演化分析。在主题挖掘方面，本文先利用 Jieba 分词工具中的 TF-IDF 模型得到热词排序表；其次，使用 wordclound 和 matplotlib 绘图工具，对舆情主题变化进行可视化展示；最后，通过 HDP-vMF 模型将热词进行主题聚合，用于研究舆情期间微博用户讨论的主题。在情感演化分析方面，本文基于 BERT 模型进行实验，探究舆情全生命周期中网民的情感倾向性与演化过程。同时，为验证 BERT 模型情感分类的优越性能，本文选取了 6 种情感分类模型 KNN、SVM、XGBoost、DNN、LSTM、Bi-LSTM 进行对照实验。

四是结合整个舆情发展中的关键时间点、关键实事进行模型验证，探究本次舆情的发展规律。从而为媒体人和政府掌控舆情走势，了解微博用户情绪变化和关注话题点提供理论及实践支撑。

1.3.2 研究方法

（1）文献研究法

针对突发公共安全事件引发的网络舆情的问题，为了有效挖掘微博用户的关注话题和情感演化过程，通过对中国知网、维普、万方、SCI 等数据库，以“主题挖掘”、“在线评论”、“情感演化”等关键词进行检索，阅读与研究大量国内外相关文献，梳理总结其研究方法及相关理论。一是探究现有学者在此领域常用的算法及模型，分析其薄弱环节与可改进之处，进而制定本文的模型框架。二是学习如何界定网络舆情发展的不同阶段，从而为本文的舆情阶段划分提供理论依据。三是梳理相关文献的实验研究架构，为本文整理出一个清晰的脉络奠定基础。

（2）案例分析法

案例分析方法也称个案分析法，相关研究学者通常会选取有代表性的的案例进行研究分析，从而得出该类事件或现象的普遍发展规律，并为后续研究者提供类似问题的解决方法。由于本文构建了一种新的混合模型，故采用案例分析法不仅是为了通过某一个典型案例预判某一类事件的走势，还希望通过该案例的实际发展趋势验证本文所提模型的准确度与可靠性，从而为后续学者及舆情监管人带来实际价值。因此，本文选取了 2021 年 7 月 19 日发生的“河南暴雨”事件为案例，通过收集数据，训练模型，探究本次事件中微博用户的关注话题和情感演化过程，并提出相应的解决策略。

（3）定量分析法

为了挖掘海量评论数据中网友讨论的热点主题及情感值，本文依据爬虫工具抓取的 120791 条相关微博评论数据，建立基于 HDP-vMF 的主题模型与 BERT 情感分析模型的混合模型，通过不断调整模型参数优化模型效果。其中，为了验证 BERT 情感分类模型的分类效果，本文选取了另外六种情感模型进行对照试验。最终，利用该混合模型计算出研究对象的各项主题聚类指标及情感数值后，结合趋势分析法得出整个舆情周期中微博用户的情感走势。

1.4 研究框架及创新点

1.4.1 研究框架

本文以“河南暴雨”事件引发的网络舆情为例，对微博评论数据进行主题挖掘并且构建分类模型进行情感分类，探究微博用户对此事件的关注主题和情感演变规律的同时，验证本模型的优越性。本文研究框架安排如下所示：

第一章为绪论。主要围绕以下四个方面具体展开：一是详细陈述与本文研究方向相关的背景及研究意义。二是国内外研究现状，主要分析主题模型研究和情感分析技术两方面的内容。三是借助技术路线图与流程分析简单阐明本文的研究内容、研究方法。四是概述论文的整体结构框架，并说明本文的创新点。

第二章为相关概念及理论基础。首先是关于本文研究对象的概念介绍，本文主要研究的是突发公共安全事件引发的网络舆情，故先介绍网络舆情的概念，以此便于与普通舆情做出区分。同时，针对该类网络舆情的特有属性及现有的舆情监测技术进行分析，为本文后续研究提供一定的启发。其次，在情感分析模型介绍部分，简单阐述本文分析情感模型的概念及现有情感分析模型的种类。由于本文所用模型 BERT 属于深度学习领域，因此只着重介绍基于深度学习的情感分析模型及算法。最后，对文本挖掘技术进行详细结束，并对比分析文本挖掘与数据挖掘的区别。

第三章为 BERT-HDP-vMF 混合模型构建过程。主要包括 DP-vMF 模型构建、HDP-vMF 模型构建、BERT 模型构建与混合模型搭建。混合模型的搭建过程主要为：做好数据预处理与词向量构建工作后，通过不断调试参数并反复试验，将用于情感分析的 BERT 模型和文本挖掘的 HDP-vMF 模型更好的拟合在一起，最终提高混合模型的性能。

第四章为以新浪微博平台的“河南暴雨”话题为例进行实证分析。一是通过爬虫工具爬取有关该话题的微博评论数据，利用正则表达式处理带有大量噪声的原始数据，借助 jieba 分词对处理好的数据进行分词，用 word2vec 与 BERT 共同

构建具有情感属性的词向量，从而为后续实验提供准确有效的数据源。二是根据实验得出舆情期间用户的情感值，绘制相应的情感走势图，从而探究微博用户的情感演化过程。三是将微博用户讨论的热点词语用词云的形式展现出来，并从中选出讨论频率最高的九个词进行排序，最后使用主题模型将热点词进行聚类，获得更加切合本次舆情的主题信息。四是利用对比实验进行比对分析，证明本模型的准确性的同时，探究河南暴雨事件所暴露的问题，进而提出适宜的解决措施。

第五章为本文总结与展望。该部分主要阐述实证分析的结果与本文所使用的模型性能，并提出整个研究过程中的不足与局限点，以及研究启示和对今后相关研究工作的展望。

1.4.2 论文创新点

（1）针对网络舆情事件构建了一种混合模型，该模型不仅可以挖掘评论文本中网民的关注热点话题，还可以分析整个舆情周期内网民的情感演化过程。故本文在考虑舆情话题的主题和情感关联性的同时，又注重探究舆情发展过程中公众的情感波动情况。

（2）在词向量构造方面，本文使用 BERT 构造的动态词向量，解决了 SVD、word2vec 等传统词汇级别静态词向量在面临语境、语序等变化时，无法有效地将原始词汇之间的语义信息嵌入到词向量的问题，即：无法解决一词多义的问题（例如，“我的手机是苹果的”和“我爱吃苹果”中的“苹果”并非同一个对象）。

（3）在情感演化分析方面，本文通过使用基于双向 Transformer 网络的 BERT 模型，既通过并行运算的工作方式在显著提升模型的训练效率，又利用多头注意力机制对句子进行全局分析解读，以此提高语句情感分类准确率。

（4）在主题挖掘方面，为了解决传统主题模型只能分别提取一个评论进行分析，从而忽略各评论之间关联性的缺点，本文将狄利克雷过程扩展为层次狄利克雷过程（HDP），利用层次的性质提取多条评论之间的语义关系，共享多个聚类主题，从而更细粒度地挖掘主题。另外，将 Gaussian 分布替换为 von-Mises Fisher 分布使得所有的簇被均匀映射到一个单位超球面上，而不是鼓励簇心向原点聚拢，从而减小聚类时产生的误差。

（5）通过爬取微博平台上有关“河南暴雨”的数据进行实证分析，并且在相同的数据和实验条件下，使用 KNN、SVM、XGBoost、DNN、LSTM、Bi-LSTM 六种常用情感分析算法，验证 BERT 模型在情感分析方面的准确性。

第 2 章 相关概念及理论基础

2.1 网络舆情的相关理论

2.1.1 网络舆情的概念及特点

网络舆情属于舆情中的一类，是指以网络为载体，以事件为核心，广大网民通过社交平台或官方媒体等发表自己对某一事件的看法与观点，表达自己的情感与态度，并通过点赞、转发分享等一系列过程的集合。通常，公众所持的观点和看法有较强煽动力和倾向性。随着社交平台的飞速发展，大多数公众都倾向于使用新浪微博、微信、抖音、知乎、官方媒体、今日头条等 APP 获取最新新闻事件，并对其转发分享与发表评论，因此，这些事件经过这一系列的反复发酵后，很容易演变成重大舆情危机事件，从而不利于社会的和谐发展[50]。

虽然网络舆情和传统媒体舆情之间有许多共通之处，但随着互联网技术的高速发展，网络舆情自身鲜明的特点愈发明显[51]。其特点形成的原因主要归结为两方面，一是网络平台的技术水平，二是网络言论管理规范尚不完善。网络舆情的主要特点为：（1）信源模糊性。网络传播中的信息来源往往不够清晰，其信息源可被分为三类：一是信息内容无法追溯源头；二是道听途说；三是凭空想象，捏造一个信息虚假信息。由于大多数网民对网络上的信息没有很强的真伪辨别能力，因此，如果信息的真实性未经权威机构证实，或是传播链未被封堵，公众往往会不假思索的选择相信并传播扩散给他人，进而引起公众热议。猎奇心理在大多数人身上都存在，一般而言猎奇心理并不会造成不良后果，但在网络的作用下，可能会产生不良影响，而且在从众心理的作用下，这类信源模糊的信息可能就会被大众接受，并被信以为真。（2）传播爆炸性[52]。在传统媒体中，舆情传播路径呈现出线性传播和圈层式传播的特征（即：人群因地域、地位、兴趣等因素被划分为不同的群体，大多数情况下信息仅在相应的群体中传播），而网络舆情传播路径则是非线性的且不受圈层限制，各传播渠道之间也相互连通，如果不加控制舆情信息便会迅速蔓延。简而言之，当某一事件发生时，网民可以没有任何地域限制、时间限制的在网络平台发表个人看法与转发，经过网民不断转发后，越来越多的人也随即参与到该事件的讨论中，由此公共意见被迅速汇聚起来，声势浩大难以在短时间内平息。（3）主动隐蔽性。在社交网络平台中，传播主体模糊，公众通常借助网络虚拟昵称，在彼此不清楚对方真实身份的情况下进行交流，网络言论与应承担的社会责任难以直接挂钩，工作者信息分析难度大[53]。因此，部分居心叵测之人借此发表不当言论，制造虚假言论，负面信息在传播过程中不断发酵，从而酿成网络舆情危机。（4）网民动员性。当某一事件出现在社交平

台，经过拥有大量粉丝的博主转发后，短时间内会调动网民的积极性，网民参与事件评论表达自己观点和看法的同时，转发与分享给他人使信息内容更加丰富，进一步加大传播效果。网络舆情既可以传播信息内容本身，又能够传播网民意见、评论及发帖或跟帖量。同时，网民可以还看到自身行为对网络舆情的影响，如果网民的自身行为可以影响舆情走向，这种成就在一定程度上可能反作用于网民使其产生满足感，进而更加积极参与讨论。当不同群体对同一事物的看法不一致时，意见表达趋于情绪化，在争论的过程中会出现人身攻击、人格侮辱等极端言论造成舆情危机[54]。

2.1.2 网络舆情的分析技术及工具

当人们通过网络获取相关信息时，也会通过各种平台表达自己的观点和看法。伴随着网络媒体时代的到来，在开放的网络环境下，信息的传播速度越来越快，对网络舆情的监控分析变得越来越重要，而监控分析技术是决定网络舆情监控分析效果好坏的核心。舆情分析系统基本环节包括：信息监测、信息管理、信息服务[55]。在这三个环节中都需要用到不同的舆情分析技术。在信息监测过程中，需要相关信息的实时检测与采集；在信息监管过程中，从网站中爬取到文本数据后，需要完成情感分类、主题聚类、主题分类等工作；在信息服务过程中，所采集到的信息最终都需要被整理成可供用户直接或间接使用的相关工具，如机器生成舆情信息报告、分析现有的舆论热点话题等。总的来说，舆情分析[56]涉及语义分析技术、文本分析技术、类信息收集技术、分类与整合技术、过滤技术等。语义分析是一种通过自然语言解析数据所传达的内容或含义的方法。语义分析技术既可以分析词语级和句子级的语意，也可以结合上下文分析段落所表达的含义。语义分析技术中涉及的词法分析可以准确地分析用户输入信息的具体特征，从而对其进行搜索。语义分析技术中涉及的句法分析可以用于识别用户输入句子的句法结构，最终完成自动分析句法这一步骤。语句分析可以根据文本数据中的上下文提取有关意象、背景等其余信息，将语言内容与现实生活连接在一起。语境分析主要用于特定领域或查询用户需求等方面，而句法结构与语义信息正是 NER 模型最为关注的信息。NER 模型借助给定的阈值，利用给实体对象打分的方式标记识别的正确率。文档主题生成模型通过训练后，可以较为准确地掌握数据之间的相似性，进而自动实现主题分类和分组的目标。LDA 模型根据词袋思想抓取数据的主题和内在含义，虽然它无法理解语言文字符号等，但是可以通过反复训练学习人类日常交流或书面表达的语法规则。LDA 可以用于分析各类结构数据，如：非结构化、半结构化、结构化，且不受数据规模限制。

面对众多新媒体舆情舆论数据信息，仅靠人工监测不仅费时费力，而且管制乏力可能引发二次危机，因此通常需要借助专门的舆论监测平台。因此，从事舆情监管和预测的工作人员在掌握必备的分析技术的同时，往往也需要借助各类舆情分析软件工具，对全网舆论进行实时监测，提高舆情信息的收集与整理效率。网络舆情分析主要是从相关舆论话题的影响力、影响层面、传播动态、网民情绪[57]等几个方面入手进行分析。开展网络舆情监管工作时往往还需要分析媒介的传播情况，通常围绕舆情的传播途径、传播速度、受众群体和受影响程度四个方面展开[58]。目前，除了可以利用搜索引擎工具、爬虫软件和平台自动的检测服务功能等信息检测平台以外，还有大数据舆情监测平台。主流大数据舆情监测工具主要分为企业舆情分析系统软件、政务舆情分析系统软件和社交媒体舆情分析系统软件。识微商情监测系统更受企业青睐，该工具可以自动分析企业品牌口碑、高管负面舆情和代言人的舆情等情况。在企业发展过程中，如何提高自身产品的市场竞争力，并在行业站稳脚跟离不开竞争情报的获取，识微商情能够自动抓取行业信息，分析现有市场竞品舆情情况。鹰眼速读网系统旨在帮助政府从业人员，以更精准快捷的方式顺利开展政务舆情监管工作。政府相关部门人员通过该平台可以了解当下民生热点问题和重大社会问题，掌握舆情传播来源、各媒体的传播情况和用户转发分享动态等。社交媒体舆情分析常用的工具为鹰击早发现系统，各级政府人员可以了解当前社会中公民对某件事情的看法；追踪事件的整个传播路径和传播源头；挖掘网民话题和情感的地域分布情况。

2.2 情感分析理论与技术

2.2.1 文本情感分析的概念

文本情感分析是一种通过计算机技术自动挖掘文本数据中评论主体的情感倾向、态度和情绪状况的过程[59]。文本情感分析涉及的领域广泛，可执行各类不同的任务，例如：情感分析、意见挖掘、主观性分析等。在工业领域，“文本情感分析”术语较为常用。在学术界，“文本情感分析”和“意见挖掘”都为常用术语。无论如何，他们基本上代表了同一个研究领域。

情感表达由四个元素构成，分别是观点持有人、评价对象、极性与时间。观点持有人是指第一个提出该观点的人；评价对象是指该观点评价的人、物或事情；极性是指该观点所体现出的情感类别（如正向、中立和负向）；时间是指文本发表的时间。由于发表时间的获取难度最低，不需要通过情感分析就可以获取，因此观点持有人、评价对象、极性三元素是情感分析的主要对象。情感分类一般由正—负情感、正—负—中情感、情感分值等体系构成。文本中的情感既有显式的

又有隐式的，显式情感是指有文本中含有明确的主观情感词（如：喜欢、开心），隐式情感是指文本中不存在任何表达感情的词语，例如："这盘菜里有蟑螂"。由于隐式情感分析难度比较大，如果不联系生活实际则很难判断其情感属性，因此，现有的研究主要围绕显式情感文本开展。情感分类及情感分析应用系统是文本情感分析研究领域中的重要组成部分之一，其系统图如图 2.1 所示：

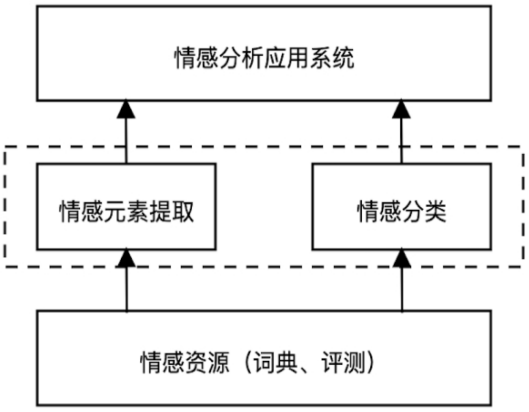

图 2.1 情感分析应用系统图

随着文本情感技术的不断发展和进步，网络中出现了许多情感分析的系统和平台，但是大多数系统都被应用于商业方面，最常见的有：商品/服务评论分析、社交平台分析、情感机器人。其中，基于日常消费品和服务的评论是传统情感分析系统的主要研究对象。谷歌公司开发的应用平台“谷歌购物”通过分析用户对产品的评价，为用户开放商品查询和比价功能，提高了用户的购物体验感。除此之外，基于高级语义和语言研究 SMI 决策平台“OpinionEQ”可以分析消费者心理，为商业组织和私人定制个性化产品分析服务。随着人们逐渐习惯于线上沟通的社交方式，以国内的微博和国外的推特为代表的社交平台成为当下使用频率最高的软件。基于此，微博和推特上产生的大量评论信息为情感挖掘工作者带去更多机遇，情感挖掘工作者通过收集分析微博/推特用户的评论数据探究公众的情感倾向和演变过程，从而为政治、娱乐等领域发展创造价值。另外，情感分析技术也被广泛应用到人机对话的智能机器人领域。其中，最为常见的是小米、百度两公司研制的智能音箱，当用户呼唤其名字″小爱同学″和″小度″时，该机器便可以通过分析用户说的话，和用户进行沟通，并且可以根据用户的指示为其提供相应的服务，例如：当用户表明自己心情不好时，智能语音音箱会根据用户心理现状安慰用户。

2.2.2 情感分析模型与方法

自然语言处理(Natural Language Processing)，简称 NLP，主要适用于计算机和人工智能领域，可以利用人类日常交流的语言和计算机交互的技术[60]。情感分析属于自然语言处理中最常见的文本分类问题，即：根据人类自然语言交流的日常习惯，自动将文本进行情感倾向分类。目前，最常用的情感分类方法有三种。第一种：基于情感词典的情感分类。基于情感词典的方法较为简单且最基础，词典词库中的语料可以加入相应研究领域的专业词，且不同词典采取的判别规则不尽相同，但基本都遵循拆解文本、抓取关键词和统计情感值的步骤。虽然情感词典便于研究者理解与操作，但这种方法通常无法考虑词与词、句子与句子之间的顺序和语法，而是将整段文本视作一个词集合，从而无法充分理解文本的语义信息。比如，有两段文本“你的表述方法很准确”，“他的表述方法不准确”，在关键词提取时，提取出两段文本的关键词均为”准确”，通过这种方法判定得出结论：这两人的表述方法都准确，显然结论是错误的。第二种：基于机器学习的分析方法。与基于词典的方法不同，机器学习算法基于其强大的推理学习能力，拟合真实世界的相关事件场景，并对后续的一些决策起到辅助作用。传统的软件程序往往用于解决特定任务和硬编码，而机器学习需要学习数据中的分布规律，从而自动完成主题挖掘、情感分类等任务。第三种：基于深度学习的分析方法。随着机器学习的发展，用于实现情感分析的机器学习算法趋于成熟后，在机器学习的基础上又衍生出深度学习。虽然深度学习属于机器学习中的一种，但其更注重深度，利用神经网络更加深层次的理解数据背后的意义[61]。首先，深度学习擅长完成文本数据处理和语义理解工作，结构更加灵活，有许多隐层节点，宽度广，能解决更复杂的问题。其次，深度学习的学习能力更强，不需要人工设计特征或规则就可以优化损失函数，达到自主学习规则的目的。最后，深度学习可以处理数量更加庞大的数据，从一定程度而言，数据量越大，深度学习的表现就越好，甚至在图像识别、自然语言处理和人脸识别等方面的表现超越人类。目前情感分析领域常用的一些深度学习神经网络主要包括多层神经网络（MLP），卷积神经网络（CNN）和长短记忆模型（LSTM）。MLP 的网络有很多隐层组成，且每个神经元都和上一层中的所有节点连接，参数量大使训练难度加大，会丢失像素间的空间信息，只接受向量输入。CNN 在很大程度上弥补了 MLP 的不足，其局部稀疏连接且参数少，接受矩阵输入，可以利用像素间空间关系，引入池化和空洞卷积。这些神经网络共同点是内部含有大量的无明显语义的参数，而这些参数基本上都是通过反向传播来更新的。虽然 CNN 和 LSTM 属于两种类型不同的模型，但二者都可以实现参数共享。这类似于传统的机器学习方法，当增强相关约束条件时，整体训练参数的可更新域越小，模型探索空间收紧，不容易发生过拟合现

象。反之，模型越大，限制越小，整体模型容易发生过拟合。这类问题通常基于一些正则化方法来解决，例如 L1，L2 正则，以及 Dropout 方法。

2.3 文本主题挖掘理论与方法

2.3.1 基于主题模型的文本挖掘介绍

文本挖掘是数据挖掘的分支，可以提取出文本背后所传达的有效信息。文本挖掘涉及多种学科，涵盖的技术种类丰富，例如：线性几何、NLP、高等数学等。基于主题模型的文本挖掘的操作步骤有六步。一是文本获取。文本的获取可以通过直接导入现有文本数据，或利用网络爬虫等技术[62]。其中，爬虫爬取的网络数据主要是以网页 HTML 形式为主，因此，需要把网络文本数据存入数据库（数据集），然后进行后续操作。二是文本预处理[63]。由于在直接爬取的数据中存在许多干扰的信息，例如：表情包、html、js 代码，注释等等，而这些干扰信息无法作为模型的输入进行训练，因此需要将其全部剔除。三是文本的语言学处理[64]。这部分首先需要对文本进行分词操作，通过某种方法将完整的句子切分成以词为单位的集合，并且分析出每个词的重要程度；其次是词性标注，标注出词语的词性，例如：形容词、名词、动词、语气助词；最后是去除停用词，每篇文章中都会出现许多类似“的”、“是”、“了”等没有实际意义的词，这些词并不能反应出文本的实际含义，故应该被去除。四是文本的数学处理-特征提取，在挖掘文本主题时，往往旨在获取既能保留文本的信息，又能反映其相对重要性的词语，那样得到的文本主题才是相对准确的。通常一个文本可能包括许多含义相近的词语，如果将其全部保留，会使词向量的纬度过高，矩阵稀疏，挖掘效果不佳，因此需要借助特征提取。五是分类聚类。对所有词语的特征或属性进行分类，然后再将特征或属性一致的词聚集到一起。其中，分类常用的方法有：贝叶斯分类，矩阵变换、SVM 等。聚类方法通常有：平面划分法、K-Means 等。六是数据可视化。通过做图工具以折线图、扇形图、词云或表格等更加直观的方式展示实验结果，便于他人理解。

2.3.2 LDA 主题模型

主题模型（Topic Model）是文本挖掘领域至关重要的工具之一，常用于文本分类，深受学术研究者和工业领域人员的高度关注。在进行文本挖掘的过程中，由于大多数数据的形式都是非结构化的，如果不借助模型算法，研究人员很难获取文本中有用的信息和知识，而主题模型可以以更加高效的方式代替传统方法，挖掘出文本数据的主题和隐藏信息。例外，主题模型还可以应用于主题聚合、抓

取指定特征等方面。LDA 是一种由词、主题、文档三层组成的文档主题生成模型，其根据概率分布生成与文档最匹配的主题。LDA 属于无监督类型的机器学习，研究者训练模型时不需要对训练数据打标签。另外，LDA 预处理过程简单且模型参数少，仅需要设置好实验最终结果所要达到的文档集和主题数，而不用通过人工标注大量的训练集。在一段文本中，任意一个主题都是通过一定概率从每个词语中生成的。假设学者需要从所有实验数据中寻找 10 个主题，LDA 的目标则是替每个文本寻找一个维度为 10 的向量作为某个主题的概率。LDA 主题模型生成过程为：Step1. 从任意一个由若干词语组成的文档中，以一定概率抽取一个主题；Step2. 从上述主题中选取一个词语；Step3. 重复上述过程，遍历每个单词（假如该文档由 2000 个词组成，则需要重复 2000 遍）。

主题模型属于典型的词袋模型（Bag of Words Model），是文本向量化的模型，其将文档视作一组含有若干个词语的集合，它只考虑带集合中的单词是否存在，而不关心语法和词语之间的先后顺序和结构信息。简而言之，被划分好的词语会被放入相应的袋子中，即使在同一个袋子，这些词语也都是彼此独立的。例如有两个句子，（1）A 从事计算机方面的研究、（2）B 从事考古学方面的研究。词袋模型的构成过程可简单描述为：首先，将两句话进行分词后分别得到两组词（1）｛A、从事、计算机、方面、的、研究｝、（2）｛B、从事、考古学、方面、的、研究｝。其次，将两组词中的重复项合并，进而将其装进同一个袋子：｛A、B、从事、计算机、考古学、方面、的、研究｝。最后，根据袋子和词语出现的次数，用 1, 0 表示原来的两句话（其中，1 表示有，0 表示无），上述两句话则为｛1,0,1,1,0,1,1,1｝、｛0,1,1,0,1,1,1,1｝。

第 3 章 BERT-HDP-vMF 混合模型构建过程

3.1 BERT 情感分类模型的构建

BERT 是 Google 公司技术人员 Devlin 等人于 2018 年 10 月提出的预训练语言模型，其网络架构是基于 Transformer 结构的 Encoder 部分，通过大规模的语料数据来为下游 NLP 任务提供丰富的语义信息。

其内部结构为基于多头注意力机制的双向 Transformer 编码器，不但可以针对性地削弱 Masked LM 任务中 mask 标记的权重，降低 mask 标记对模型训练的不利影响，还能够利用并行运算进行计算工作，显著提升模型的训练效率。同时，多头注意力机制不需要考虑字符的方向和距离问题，可以直接互通任意两个字符，解决以 RNN 为结构带来的长距离依赖问题，从而通过上下文更好地学习文本的语义表示，增强模型语义分析理解的能力。由此，BERT 与其他语言模型相比，为情感分析研究提供了更加匹配的理论算法。目前诸多学者的研究[65][66][67][68]已经初步证明使用 BERT 作为预训练模型可以大幅提升下游 NLP 任务的完成效率。

为了利用 BERT 进行情感分类研究，本文在 BERT 模型后拼接了基于神经网络的情感分类模块，整体模型结构如图 3.1 所示。

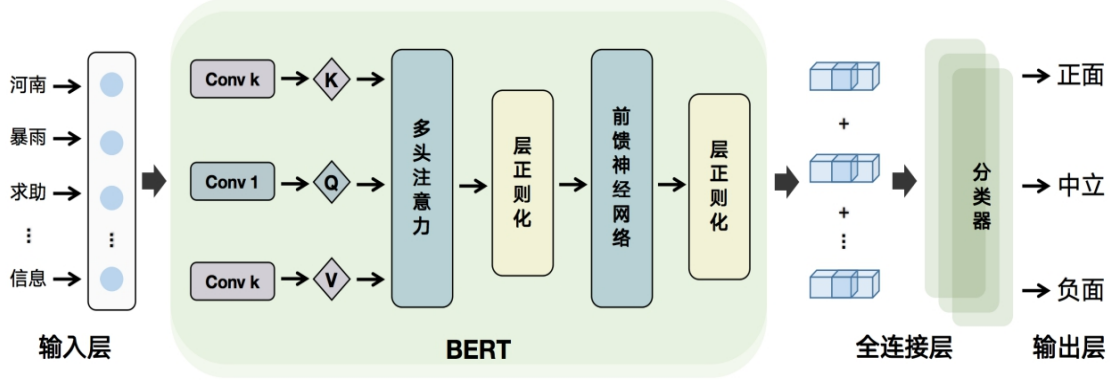

图 3.1 基于 BERT 的情感分类模型结构图

通过图 3.1 可以看到，整个情感分类模型一共分为四层，它们分别是：输入层、BERT 模型、全连接层、输出层。

第一层为输入层。由于计算机无法识别中文文本，因此在中文文本输入模型前需要按照实验者指定的要求，把文本切分割并转化为字符串，而这一过程也被称为 tokenization。具体来说，tokenization 就是先对文本进行分词，把本文分割成一个个独立的 token（即一个中文对应一个 token），然后构建词库，最后再对

每一个 token 赋予相应的 ID。总的来说，第一层输入层的主要作用是将原始文本信息转化为可用于模型训练的 embedding。首先通过将数据进行预处理操作得到词汇集$W=\{w_1,w_2,...,w_o\}$，其中o为词语集单词数，基于所得词汇构建“词汇—词汇 ID”映射表 S，通过映射表将输入的词汇转换为 ID 输入 BERT，此时词汇$w_i=\{id_1,id_2,...,id_c\}$，其中$c$为第$i$个词汇的字数。由于数据是以矩阵形式作为输入的，所以只有数据长度一致，才能构成一个矩阵。所以，为了确保输入数据长度一致，需要先计算所有词汇中的最大词汇长度m，对超出长度c的部分进行补 0 操作，此时$w_i=\ id_1,id_2,...,id_{c,...,}0_m\}$。

第二层为 BERT 模型。该层主要利用自注意力机制来获取到词与词之间的语义信息。本文在预加载基于百度百科预训练好的模型参数的基础上，通过使用原始文本对其性能进行进一步微调的方式对模型进行预训练，从而获取词语之间的语义信息。所有层都联合上下文语境进行预训练是 BERT 训练语言模型特点之一，其预训练采用了两个训练任务。其中一个任务是用来捕捉单词级特征的 MLM（Mask Language Model）任务，即随机从输入预料上遮蔽掉一定比例的字词，然后通过的上下文预测该字词，类似于一个完形填空任务。例如，将“中国的首都是北京”这句话中的“北京”或“中国”任意一个词语遮蔽掉后，让这句话变成“xx 的首都是北京”或“中国的首都是 xx”的形式供模型训练，经过一定大规模的数据训练后，模型便可以正确预测出被遮蔽的字词。另一个任务是用来捕捉句子级特征的 NSP（Next Sentence Prediction）。该任务的主要目的是判断语句 B 是否是语句 A 的下文。生成训练数据的方式是从语料库中随机抽取一批连续的两句话，其中 50%会原样保留语句，作为“IsNext”的训练样本，另外 50%的数据会将其第二句话替换成任意一句话（替换的新语句一定不是原来的第二句话），作为“NotNext”的训练样本。

第三层为由两个全连接神经网络构成的情感分类器。全连接神经网络是深度学习中最基础的网络结构之一，主要的核心作用在于将高维的向量通过矩阵乘法的方式降为低维度的向量，而网络中的权重参数用来保存高维向量到低维向量之间的映射关系，基础公式为：

$$x_1=weight*x+bias \qquad \text{式（1）}$$

其中x为输入的高维向量，$weight$为网络的权重参数，用于保存映射关系；$bias$为偏置参数，与线性方程一样，该参数主要用于让函数偏离原点，增强函数的灵活性与拟合能力。经过前两个步骤的处理后，每个词语都有其对应的向量，该向量值可以为正值、负值或是 0，通过将一段语句中所有词语进行加权平均得到的值即为句子表示向量。在得到句子的向量值后，将 BERT 的输出作为分类器的输入来进行句子情感分类，具体来说，第一个全连接神经网络的权重参数大小

为 512*256，偏差参数大小为 512*1；经过上述公式第一个全连接神经网络的输出大小为n*256，其中n为数据量；第一个全连接神经网络后连接了一个激活函数 ReLU，该函数主要作用为增强整体网络的非线性能力，从而让网络能够拟合更加复杂的数据，核心思想为当$x_1 > 0$时输出x_1，当$x_1 < 0$时输出 0，这样的一个性质也同样避免了网络在反向传播时容易产生梯度爆炸，消失的问题。第二个全连接神经网络的权重参数大小为 256*3，偏差参数大小为 256*1，将x_1输入第二个网络可以得到输出大小为n*3 大小的向量，其中维度设置为3主要是因为我们预先设定了情感极性有三种。最后n*3 大小的向量被输入 softmax 函数，该函数主要作用在于得到输入向量的概率分布，即 3 个维度的数据都小于 1，且和为 1。如果输出概率值处于正向那一维度的概率最大，则这条语句的情感值即为正向；如果输出概率值处于中性那一维度的概率最大，则这条语句的情感值即为中性；如果输出概率值处于负向那一维度的概率最大，则这条语句的情感值即为负向。

第四层为输出层，主要用于输出每个句子的情感类别。通过利用原始文本对整个模型的训练，分析模型不仅能得到原始文本所隐含的情感类别，还能够获得更加符合原始文本语义信息的词向量，这部分词向量将为本文在主题模型构建方面奠定坚实的数据基础。

3.2 HDP-vMF 主题模型的构建

狄利克雷过程（DP）[69]是一种应用于非参数贝叶斯模型中的随机过程，其具体内容为：假设G_0是测度空间θ上的随机概率分布，参数α_0是正实数，空间θ上的概率分布G如果满足以下条件：

对测度空间θ的任意一个有限划分$A_1...A_r$，均有以下关系：

$$(G(A_1),...,G(A_r)) \sim Dir(\alpha_0 G_0(A_1),...,\alpha_0 G_0(A_r)) \qquad \text{式（2）}$$

则可以得出G服从以基分布G_0和参数α_0组成的狄利克雷过程：

$$G \sim DP(\alpha_0, H) \qquad \text{式（3）}$$

狄利克雷过程被广泛应用于大规模语料下的主题提取（LDA），通过将文档的主题以概率分布的形式来进行聚合分析，其建模流程见图 3.2。给定一个包含n条句子的文档D，$D = \{S_1, S_2, ..., S_n\}$；$W = \{w_1, w_2, ..., w_o\}$。

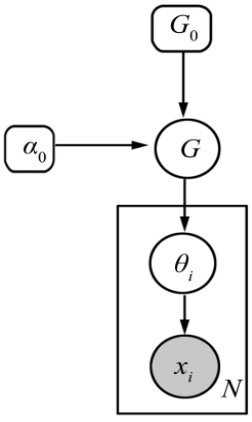

图 3.2 DP 结构图

建模过程可解析为：

Step1：假设文档 D 的先验分布为 $P(D)$；

Step2：从以 α_0 和 G_0 为参数的狄利克雷过程中采样主题分布 G；

Step3：从主题分布 G 中取样第 i 个主题 θ_i；

Step4：给定参数 θ，并从中生成主题 θ_i 的分布 $\varnothing_i$；

Step5：从 $\varnothing_i$ 中生成词语 w_i。

但传统的基于狄利克雷过程的主题提取算法只能对单组数据进行聚类，无法对多文档进行并行化的主题提取，从而丢失了文档间的语义信息，并且需要花费大量的训练时间。在微博评论数据中，每一条评论理应被单独作为一组数据，多组数据之间共享相同的主题。因此，为了更好地利用多条评论之间的语义关系，共享多个聚类主题，本文将狄利克雷过程扩展为层次狄利克雷过程（HDP）[70]。相比于单一的 DP，HDP 通过同时对多组数据进行并行化聚类分析，可以有效地获取到多组数据之间的共享关系，挖掘出更加合理的主题信息，其有向图如图 3.3 所示。

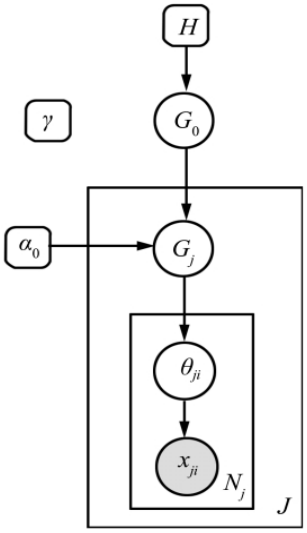

图 3.3 HDP 有向图

HDP 建模过程可解析为：

Step1：假设文档 D_i 的先验分布为 $P(D_i)$；

Step2：从以 γ 和 H 为参数的狄利克雷过程中采样共享分布 G_0；

Step3：从以 α_0 和 G_0 为参数的狄利克雷过程中采样主题分布 G_j；

Step4：从主题分布 G_j 中取样第 i 个主题 θ_{ji}；

Step5：给定参数 θ，并从中生成主题 θ_{ji} 的分布 $\varnothing_{ji}$；

Step6：从 $\varnothing_{ji}$ 中生成词语 w_{ji}。

为了使得模型能够自动确定主题数，本文模型通过使用 stick-breaking 构造法对 HDP 主题模型进行建模[71]。根据 HDP 建模过程，stick-breaking 构造法被分别用于共享分布与主题分布。首先共享分布的狄利克雷过程可表示为：

$$\begin{aligned} &\beta_k^{'} \sim Beta(1,\gamma) \quad \alpha_k \\ &\beta_k = \beta_k^{'}\prod_{m=1}^{k-1}\left(1-\beta_m^{'}\right) \quad G_0 = \sum_{k=1}^{\infty}\beta_k\delta_{ak} \end{aligned} \qquad \text{式（4）}$$

公式 4 中的 β_k 为从以γ为参数的 *Beta* 分布中采样出的值，其范围为 0-1]；β 是混合模型的混合系数，其范围为 0-1]并且满足：

$$\sum_{k=1}\beta_k = 1 \qquad \text{式（5）}$$

主题分布 G_j 的狄利克雷过程与第一层类似，通过将基分布 H 和参数 γ 替换为 G_0 和 $\imath$，可表示为：

$$\begin{aligned} &\pi^{'}_{jt} \sim Beta(1,\tau) \\ &\pi_{jt} = \pi^{'}_{jt}\prod^{t-1}\left(1-\pi^{'}_{\mathrm{jm}}\right) \end{aligned} \qquad \text{式（6）}$$

最终，根据 stick-breaking 构造法对 HDP 主题模型进行建模分析。此外，为了更好地拟合文本数据，本文将传统的高斯混合先验替换为 von-MisesFisher 混合先验，von-MisesFisher 分布是基于方向数据的分布，其定义为：数据在空间中的方向其重要程度大于数据值，典型的方向数据满足 L_2 范式：

$$\| x \|^2 = 1 \qquad \text{式（7）}$$

相比于高斯分布，von-MisesFisher 分布在高维情况下会将数据均匀地分布在超球面上，而高斯分布会促使所有簇心向原点聚拢，这并不利于数据的聚类分析。现有的研究[72][73][74][75]也充分证明了使用 von-MisesFisher 分布作为先验分布能够极大地提升聚类性能。在本文所提出的模型中，假设所有主题服从 von-Mises Fisher 分布，该分布将文本向量定义为方向数据并且投射至高维的单位超球面上，通过对超球面上的原始数据进行划分来进一步地获取到主题信息。图 3.4 和图 3.5 分别展示了模拟数据在超球面以及欧式空间上的情况，该数据从四组拥有

不同参数的 vMF 和高斯分布采样所得，基于 vMF 采样得到的簇间距大，且簇内紧密，而基于高斯采样得到的簇倾向于向原点聚集如图 3.5 所示。对于主题提取来说，主要的目的在于最大化不同主题词之间的相似度，且最小化相同主题内词的相似度，vMF 在性质上更适用于主题提取。

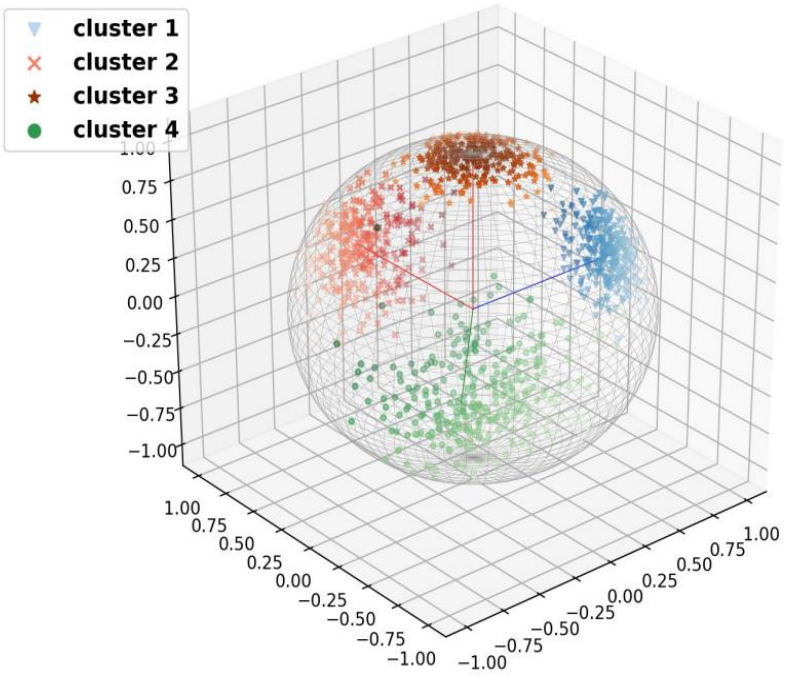

图 3.4 模拟数据在超球面的可视化映射图

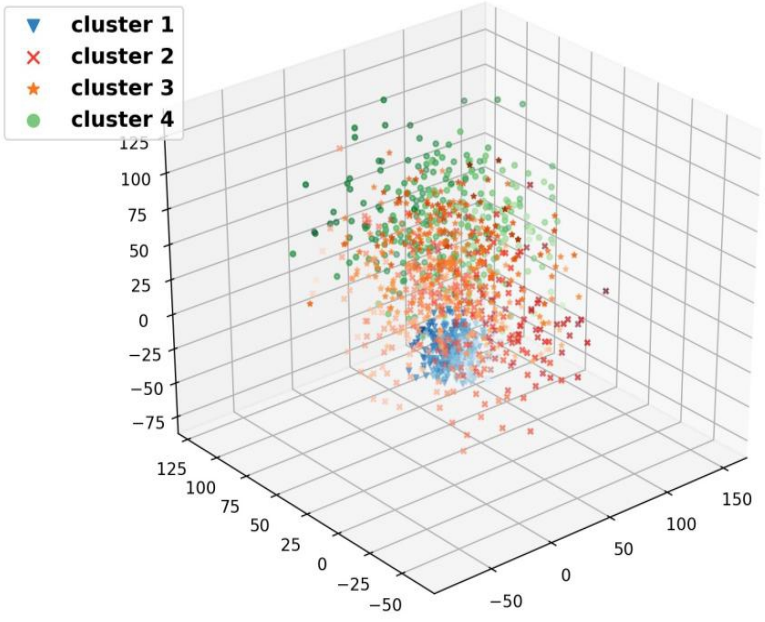

图 3.5 模拟数据在欧式空间的可视化映射图

3.3 BERT-HDP-vMF 混合模型的构建

为了对复杂语境下的大规模文本数据进行情感分析和主题提取，本文对单阶段的 BERT 与 HDP-vMF 模型进行过程优化，提出了一种 BERT-HDP-vMF 混合模型，整体流程如图 3.6 所示。

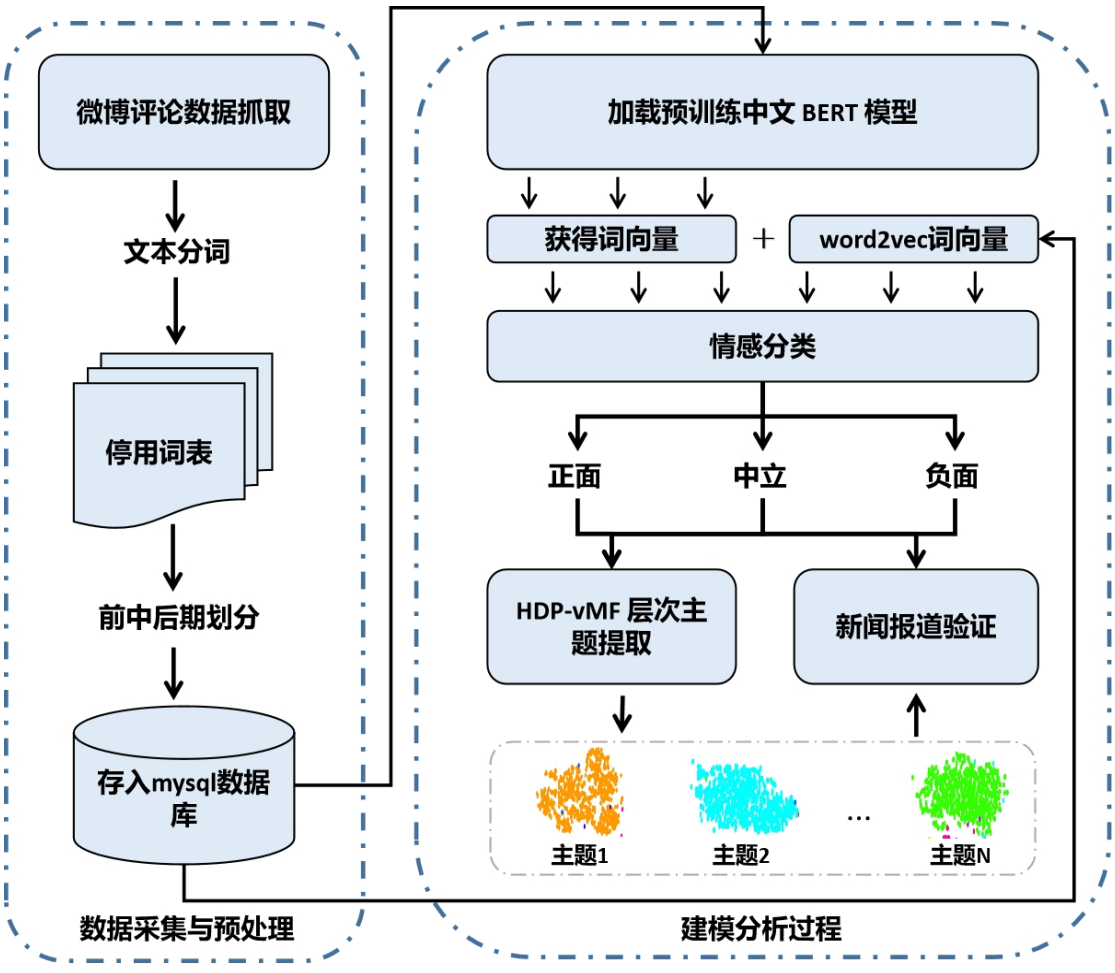

图 3.6 基于 HDP-vMF 和 BERT 模型的主题挖掘和情感分析算法流程

该混合模型可被描述为以下四个步骤：

Step1：大规模语料下的文本数据预处理；

Step2：训练 BERT 情感分类器；

Step3：构建文本词向量；

Step4：HDP-vMF 模型主题聚合。

3.3.1 大规模语料下的文本预处理

在网络平台中，任何新闻事件的传播者和被传播者都呈现出多元化和复杂化的特征，所有人都可以发表对该事件的评论，例如，记者、编辑人等专业的媒体人士，以及许多以自由方式随机表达的群众。由于这些大众群体的文化水平各不

相同，语言表达习惯也不一致，因此书面语句不一定准确、规范，甚至在语句中还会夹杂着一些英文字符、网络表情和符号等。虽然这些表情符号并不影响他人阅读，但是对于计算机模型来说，这些信息却属于干扰信息，其无法作为模型的输入进行训练。而爬虫工具在爬取网民评论数据时，只会原封不动地爬取这些原始评论，并不负责对其进行处理将其转化为模型可以识别的数据，因此，为了获取到模型可以识别的数据，本文需要对原始数据进行数据清洗，清洗操作主要包括两步：

step1.删除表情符号以及标点符号；

step2.获取每条数据的词语集。

在步骤（1）中，本文通过使用正则表达式进行特殊字符匹配的方式删除表情符号以及标点符号。在此之前，先简单介绍一下正则表达式，它是一种用来匹配字符串（我们可以将字符理解成一个字或是词语，而字符串则为一串由字符组成的句子）的有效方式，可以实现对字符串的删除、增添、查询、修改等工作。例如，当用户输入邮箱账号后，系统需要判断该邮箱格式是否正确的这一过程往往需要好几步才能完成。另外，由于不同格式邮箱的判断方法不同导致代码无法复用，而正则表达式就是被设计以某种规则来快速匹配某种类型子串的方法。在本文中，为了确保给模型提供更为精准且低噪的微博评论数据，需要数据中没有标点符号和网络表情等特殊字符，具体操作如下：首先，读取储存在 mysql 中的第一条原始评论数据；其次，使用 replace 语句将字符串中除英文字母和汉字、数字等字符用空字符替换，便可以完成特殊字符的剔除工作；最后，将剔除后的数据再存回数据库。循环重复上述三个步骤，直至数据库中所有原始评论数据均被处理完毕后，便可得到全新的数据集。

在步骤（2）中，本文对原始数据进行分词以及去除停用词操作便可得到每条数据的词语集。中文句子与英文不同，词语之间边界模糊没有间隔，因此，为了让计算机更容易理解文本，通常中文信息处理的第一步是中文分词。中文分词是将一段表述里的词汇进行分解，使计算机通过训练先了解学习每个独立词汇的含义，进而实现对整个文本的理解。在众多分词工具中，本文选用 Python 中文分词库中的 jieba 分词进行分词操作，因为 jieba 工具擅长中文分词，而且除了原始的切词外, jieba 能够更加细致的切分文本，从而为后续工作奠定坚实基础。在分词方式方面，jieba 分词有两种分词模式，分别为全模式和精确模式。在精确模式下，被划分出的词语之间并不会产生歧义。基于此，本文所有评论文本的分词工作均在精确分词模式下完成。