

# 基于自然语言处理技术剖析中国文学创作的情感特征

刘鹤

商丘职业技术学院, 河南 商丘 476000

**摘要:** 本文以中国文学作品为研究对象, 以自然语言处理为主要研究手段, 对其进行研究。本项目以中国文学为例, 以自然语言处理技术为基础, 以自然语言处理为主要工具, 对中国文学作品中的情绪表现形式进行了深入的研究。研究显示, 自然语言处理方法可以有效地对文本中的情绪进行识别与归类, 为中国文学的深度解读开辟了一条全新的道路。

**关键词:** 自然语言处理; 中国文学; 情感分析; 文本挖掘

## Analysis of Emotional Features in Chinese Literary Creation Based on Natural Language Processing Technology

Liu, He

College polytechnic, Shangqiu, Henan, 476000, China

**Abstract:** This article focuses on Chinese literary works as the subject of study, utilizing natural language processing (NLP) as the primary analytical tool. Taking Chinese literature as an illustrative case, the project conducts an extensive investigation into the emotional expressions embedded within these literary texts, building upon the foundation of NLP techniques. The findings reveal that NLP methods can effectively recognize and categorize emotions within the text, thereby paving a novel avenue for the profound interpretation of Chinese literature.

**Keywords:** Natural Language Processing; Chinese literature; Emotional analysis; Text mining

DOI: 10.62639/sspis09.20240102

中国文学作为一朵灿烂的奇葩, 在几千年的发展过程中, 积淀了深厚的感情, 积淀了深厚的哲理。在此基础上, 提出了一种新的研究方法, 即将该方法应用于文学作品中, 并对其进行情感分析。因此, 本项目拟以自然语言处理理论为基础, 探索自然语言处理在中国文学作品中的情感特点, 为文学作品的研究提供新的思路和方法。

### 一、自然语言处理技术概述

自然语言处理 (Natural Language Processing, NLP) 作为计算机科学和人工智能领域的重要分支, 它的核心目标是使计算机能够理解、解释和生成人类自然语言。这一领域的研究不仅包括语言的基础语法结构分析, 还涵盖了语义理解、语境分析、情感分析等更为复杂的内容。随着近年来机器学习和深度学习技术的飞速发展, NLP 在多个方面取得了显著的进展。例如, 文本挖掘技术使得从海量文本数据中提取有价值信息成为可能, 情感分析技术则让计算机能够识别和分析文本中的主观情感倾向。这些技术的发展, 极大地扩展了计算机处理语言的能力, 使其在信息检索、机器翻译、语音识别等领域得到广泛应用。深度学习在 NLP 中的应用尤为显著, 通过构建复杂的神经网络模型, 计算机能够更有效地学习语言的多层次特征, 从而更准确地进行语言理解和生成<sup>[2]</sup>。例如, 使用 BERT 和 GPT 这一类基于 Transformer 架构的模型, 能够在文本分类、问题

回答系统、自然语言生成等方面展现巨大的潜力。此外, 随着大数据时代的到来, 数据驱动的 NLP 研究成为主流, 通过分析海量的语料库, 计算机能够不断优化其语言处理模型, 从而提高处理效果。

### 二、运用自然语言处理技术对中国文学作品的情绪特点进行解析的具体过程

#### (一) 文本收集

文本收集是对文本进行挖掘的基础与出发点, 它旨在对与研究对象密切相关的文本信息进行有效地整合与组织。在文学研究领域, 这一步往往需要对文献进行鉴别与提取, 以保证其对文本进行深度分析。比如, 如果要对某一种文学流派或某一时代的作品进行研究, 就需要从数字图书馆和网上数据库中搜集具有代表性的著作<sup>[3]</sup>。同时, 为了保证数据来源的合法性, 必须保证所有来源都遵守著作权及相关条款, 以防止侵权行为发生。

#### (二) 材料清理

材料清理对文本数据进行处理与处理的一个重要环节, 其目的在于剔除噪音、不相关的信息, 为后续的文本分析打下良好的基础。在对文献等文字资料进行处理的过程中, 往往存在着诸如标点符号、特殊字符、HTML 标记等非规则、冗余信息等非规则、冗余信息等多种非规则冗余信息, 因此, 本文提出了一种基于语义分析的非规

(稿件编号: IS-24-2-1009)

**作者简介:** 刘鹤 (1984-07), 性别女, 民族: 汉, 籍贯河南省商丘市, 学历硕士, 商丘职业技术学院 + 讲师, 研究方向: 现当代文学。

则冗余信息提取方法。若不对其进行有效的处理, 将严重影响后续分析的精度与效率。

首先, 可以使用正则表达式, 这是一种强大的文本处理工具, 用于识别和替换文本中的特定模式。例如, 在 Python 中, 可以利用 re 模块提供的正则表达式功能, 去除或替换文本中的 HTML 标签、特殊字符、数字等<sup>[4]</sup>。例如, 通过编写相应的正则表达式, 可以快速识别和删除 HTML 文档中的标签结构, 只保留文本内容。同时, 还可以利用正则表达式处理文本中的空格、换行符等格式问题, 使得文本呈现更加整洁的形态。

其次, 数据清洗还可能涉及到文本的语言预处理, 如分词、去除停用词等。这些步骤有助于提取出文本中的关键信息, 为后续的词频分析、主题建模等提供准备。例如, 可以使用自然语言处理库如 NLTK 或 spaCy 进行文本的分词处理, 将连续的文本切分为独立的单词或短语。同时, 去除停用词(如“的”、“是”等常见但信息量小的词汇)也是一个重要环节, 这有助于减少数据的噪声, 突出文本中的关键信息。

表 1 数据清洗步骤及其相关技术工具

步骤	描述	技术或工具
正则表达式处理	用于识别和替换文本中的特定模式。	Python 的 re 模块
去除 HTML 标签	删除文档中的 HTML 标签结构, 保留文本内容。	正则表达式
格式处理	处理文本中的空格、换行符等格式问题。	正则表达式
分词处理	将连续文本切分为单词或短语。	NLTK、spaCy 等 NLP 库
去除停用词	删除常见但信息量小的词汇。	NLTK、spaCy 等 NLP 库

### (三) 文本预处理

文本预处理是文本分析中至关重要的一步, 它旨在将原始文本转化为更适合后续分析的格式。这一过程包含多个关键步骤, 其中最首要的是分词 (Tokenization)。在分词过程中, 文本被拆分为更小的单位, 如单词或短语。例如, 英文文本中的句子 "The quick brown fox jumps over the lazy dog" 会被分割成 "The", "quick", "brown", "fox", "jumps", "over", "the", "lazy", "dog" 等单词。这个步骤对于理解文本结构和含义至关重要。接下来是去除停用词 (Stop Words Removal)。停用词是文本中常见但对于理解文本的意义并不重要的词, 例如 "the", "is", "at" 等。这些词在文本中频繁出现, 但对于捕捉文本的主要意义并无显著帮助。去除这些词可以有效减少数据维度和提高处理效率。词干提取 (Stemming) 和词形还原 (Lemmatization) 是接下来的关键步骤, 这两种技术都旨在将词汇还原到其基本形式。例如, 词干提取可能将 "running"、"runs" 和 "ran" 都简化为 "run"。而词形还原则会更加细致地考虑词汇在句子中的语用使用, 将 "better" 还原为 "good"。这些技术通过减少单词的变体数量, 帮助模型更有效地学习和理解文本。最后, 在词性标注 (Part-of-Speech Tagging) 这一预处理步骤中, 将每个单词标记为名词、动词、形容词等,

有助于理解单词在句子中的作用和意义。这些预处理步骤共同构成了文本挖掘的基础, 为深入分析和理解文本内容奠定了基石。

### (四) 特征提取

完成文本预处理后, 下一步是将处理过的文本转化为机器学习算法可以理解的数值格式, 这一步骤称为特征提取。在自然语言处理中, 常用的特征提取方法包括词袋模型 (Bag of Words, BoW) 和 TF-IDF (Term Frequency-Inverse Document Frequency)。

词袋模型是一种简单直观的方法, 它将文本转换为单词出现次数的向量。例如, 考虑一个简单的文本集合: ["The cat sat on the mat", "The dog sat on the log"]。应用词袋模型, 我们首先构建一个词汇表: {"The", "cat", "sat", "on", "the", "mat", "dog", "log"}。然后将每个文本转换为该词汇表中单词的出现频率向量。这种方法简单但有效, 尤其是在处理大规模文本数据时。

然而, 词袋模型忽略了单词出现的顺序, 并且对高频词汇赋予了过多的权重。为解决这一问题, TF-IDF 模型被提出。这一模型考虑了词汇在特定文档中的频率 (TF) 和在整个文档集中的稀有程度 (IDF)。举例来说, 如果某个词在一个文档中频繁出现, 但在其他文档中很少出现, 那么这个词对于特定文档的

重要性可能很高。TF-IDF 通过下面的公式计算每个词的重要性:

$$TF - IDF(t, d) = TF(t, d) \times IDF(t)$$

其中,  $TF(t, d)$  是词  $t$  在文档  $d$  中的频率,  $IDF(t, d)$  是逆文档频率, 计算公式为:

$$IDF(t) = \log\left(\frac{\text{总文档数}}{\text{包含词 } t \text{ 的文档数} + 1}\right)$$

通过这种方式, TF-IDF 能够减少常见词汇的影响力, 并增强对特定文档有重要意义的词汇的重视。这使得 TF-IDF 在很多文本挖掘任务中成为一种有效的特征提取方法。

### (五) 探索性数据分析

探索性数据分析 (EDA) 是对数据进行初步检查的过程, 它有助于识别数据中的模式、异常、结构和关键变量之间的关系。在文本数据处理中, EDA 通常涉及多种技术, 如词频分析、情感分析和主题建模, 以提取有价值的洞见。

以 LDA (Latent Dirichlet Allocation) 模型为例, 这是一种常用的主题建模技术, 它可以从大量文本数据中发现潜在的主题<sup>[5]</sup>。LDA 是一种无监督的机器学习算法, 它假设文档是由一组主题的混合构成, 而每个主题则由一组词的分布构成。在应用 LDA 时, 首先需要确定主题的数量, 这通常是一个经验决定的参数。然后, 算法会随机分配每个文档中的每个词到一个主题中, 随后迭代地调整这些分配, 以最大化词的共现概率。

对于情感分析, 其目的是识别和提取文本数据中的情感倾向。这通常涉及到自然语言处理技术, 将文本数据转化为能够表示情感的特征向量。情感分析可以是二元的 (例如, 正面或负面),

也可以是更复杂的情感分类(例如,快乐、悲伤、愤怒等)。情感分析通常使用预先训练好的模型,如递归神经网络(RNN)或卷积神经网络(CNN),这些模型能够从文本数据中提取情感相关的特征。

在探索性数据分析的过程中,词频分析也是一个重要的步骤,它涉及统计文本中各个词的出现频率。这有助于识别文本中的关键词或短语,进而揭示数据的主要主题或关注点。词频分析通常是通过创建词袋模型来完成的,即将文本转换为词的向量,每个向量的元素代表相应词在文本中的出现频次。

### 三、案例分析

在下文中,将通过两个具体的案例,展示NLP技术在文学作品情感分析中的应用,以及如何通过技术手段对文学作品进行新的解读和认识。

本项目以《红楼梦》为研究对象,通过词频分析、情绪标记等手段对其进行研究。首先,对《红楼梦》中出现频率最高的“悲剧”、“爱情”、“命运”等关键词进行了词频分析,并对其进行了定量分析。我们可以建立一套《红楼梦》字段加工流程,并对每一字词在文中的使用频率进行统计。在此基础上,利用TF-IDF(词频-反文档频度)等方法,对语料中的词语进行重要度分析。在文章的结尾,我们看到了“悲剧”、“爱情”、“命运”这三个词在小说中频繁出现的现象,这充分显示了小说中所蕴含的深刻的悲剧意味,同时也体现了作家对人物命运的探索。其次,通过使用情绪分析工具,我们可以对文本中的情绪进行评价。在此基础上,利用情绪字典对每一句话或一段的情绪进行标注,并将其归类为积极的、消极的和中性的情绪。通过情绪评分等手段,可以对文本情绪的总趋势进行定量表达。比如,给每一个情绪词语赋予一个加权,消极词语(比如“悲伤”、“失望”)是消极的,而对于“喜悦”、“欢乐”之类的积极词语是积极的。最后,将所有情绪得分加总并平均,得到总体情绪得分。通过对《红楼梦》文本的剖析,可以发现其浓重的哀伤与乡愁色彩十分明显,折射出作家对现实生活的无奈与留恋。

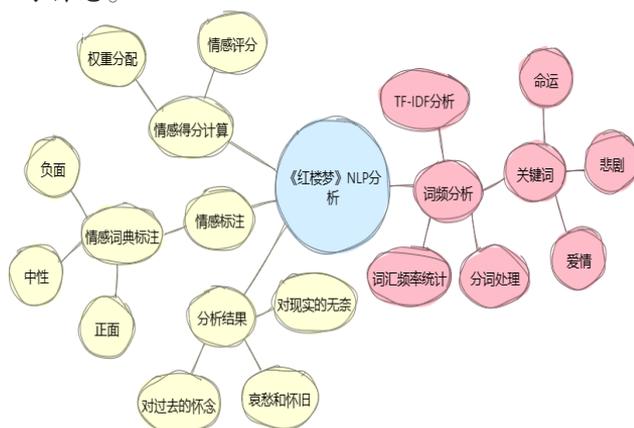


图1 NLP技术在《红楼梦》情感特征分析中的应用

对《诗经》的篇章,我们也运用了相似的自然语言处理方法进行了研究。《诗经》包含了情爱、农耕、战争等诸多领域,对其进行情感剖析,可以发现它所具有的特殊的情感特质与文化价值。然而,由于《诗经》中使用的文字与现代汉语有很大的不同,所以有必要为其专门设计一套适合该语言特点的情绪字典。在对《诗经》文本进行细致的情绪注释时,我们可以看到,诗歌中所表现出来的各种情绪情绪,特别是乐观、生气等情绪表现得特别明显。它既是古人对生命的眷恋,又是对社会不公正的控诉,又是一种复杂的情绪与价值观念的折射。比如,将《诗经》各篇章加以归类,“国风”一章侧重于情爱与民俗,“大雅”、“小雅”则多侧重于社会、政治等方面的内容,这样的归类分析,将有助于我们更好地把握《诗经》各篇章所蕴含的情感特色与文化意蕴,从而更好地展现出其所蕴涵的思想与文化特质。

对当代文学的研究表明,自然语言处理技术在各个时代都具有普遍的普遍性。当代文学往往蕴含着更复杂、更隐蔽的情绪,自然语言处理技术可以利用算法模型对其进行更深层次的挖掘。比如,自然语言处理可以从符号化的角度来理解人类的情绪,比如快乐、悲伤等,同时也可以深入到人类的内心世界,比如对自由的向往,对社会现实的思考等等。

### 四、结语

总之,自然语言处理在中国文学文本中表现出来的情绪特性具有很强的应用价值,这对我们深入了解文学作品的情绪表达有很大的帮助,同时也为文学研究提供了新的研究角度和思路。将来,随着科技的进步,自然语言处理将会在文学研究中扮演越来越重要的角色。

#### 参考文献:

- [1] 文采. 自然语言处理技术背景下的学术期刊智能化发展策略研究[J]. 传播与版权, 2024, (02): 51-54.
- [2] 雷晨. 产业化浪潮来袭: A股公司技术层加速迭代升级[N]. 21世纪经济报道, 2024-01-16(009).
- [3] 马甜, 张国梁, 郭晓军. 基于深度学习的自然语言处理攻防研究综述[J]. 中阿科技论坛(中英文), 2024, (01): 98-102.
- [4] 李希朋, 周云. 基于自然语言处理的情报文本抽取与分析技术研究[J]. 无线互联科技, 2023, 20(24): 157-159.
- [5] 龚倩. 自然语言处理技术在特检文本中的应用前景分析[J]. 西部特种设备, 2023, 6(06): 49-52.