面向大规模文本数据的自然语言处理与情感分析系统设计

郭建发

北京双动新力信息科学研究院；北京市通州区；101100；

摘要：随着互联网的发展，以微博、抖音、知乎等为代表的社交媒体成为人们表达自我、传递情感的主要阵地。社交媒体上用户发布的大量文本数据，为挖掘用户对某个产品的评价提供了丰富的数据源。近年来，针对社交媒体上用户对某产品的评价，各大科技公司和企业纷纷投入大量资源进行情感分析。但是，在大规模文本数据中，对其进行情感分析仍存在较大的困难。

关键词：大规模文本数据；自然语言处理；情感分析系统设计

DOI：10.69979/3060-8767.25.05.078

引言

近年来，社交媒体发展迅速，成为人们表达情感的主要平台。通过社交媒体用户对产品的评价可以获取用户对产品的态度，帮助企业更好地了解用户需求和潜在需求，从而提升产品和服务质量。针对社交媒体中用户对某产品的评价，国内外相关学者和企业都进行了深入研究。但是，在面对海量文本数据时，如何对其进行情感分析仍是一个具有挑战性的问题。本文以微博平台为例，采用大规模文本数据处理技术和自然语言处理技术，设计并实现了一套面向大规模文本数据的自然语言处理与情感分析系统，为企业进行社交媒体中用户评价的情感分析提供了一种新思路。

1 大规模文本数据处理技术

1.1 大规模文本数据特点

（1）数据规模大。随着互联网技术的不断发展，网络中的文本数据呈现出爆发式增长。（2）结构复杂。互联网文本数据具有非结构化的特点，多以网页形式呈现，文本数据的语义理解难度大。（3）多源异构。互联网文本数据的来源主要是网站和社交媒体，以及商业网站、电子邮件、博客等其他来源，且来源多样，格式也各不相同，这就给自然语言处理带来了很大的挑战。（4）动态变化。互联网文本数据具有时间和空间的动态性，随着时间、地点、用户等信息变化，文本数据会产生新的特征。同时，互联网文本数据还具有多源异构、动态变化、海量存储等特点[1]。

1.2 文本数据预处理

①中文分词，是将文本按照一定的规律组织成词语，为后续的文本处理做准备。中文分词主要有词性标注和词干提取两种方法。②词性标注，中文分词后需要进行词性标注，如果词语之间有某种联系，可以在词性标注的基础上进行扩展。③去除停用词：在中文文本中，会存在很多与主题无关的词语，例如“汽车”“足球”等词语，这类词语在文本中会被识别出来并去除。④去除标点符号有逗号、顿号、句号等多种类型，在文本处理中也要根据实际情况进行选择。⑤分词和词性标注是对文本进行预处理的关键步骤。对于中文文本来说，分词的主要目的是将中文分词成一个个单词[2]。

1.3 文本数据表示与特征提取

文本数据的表示方式主要有基于词袋模型、词形还原和基于向量空间模型等。在中文处理中，常见的文本表示方法包括：词袋模型、TF-IDF、依存句法分析等。其中词袋模型是文本数据的常用表示方法，即用一个包含所有单词的词典来表示文本数据，每个单词都有自己的位置，且与其他单词之间存在一种相关关系。TF-IDF则是通过计算每个单词与所有文档（包括已标注和未标注）之间的相似度，并将其作为权重来计算每个文档对整个文档集的重要性。基于向量空间模型是将文本数据以矩阵形式表示，并可以通过矩阵运算来实现对文本数据的降维处理。

1.4 文本数据存储与管理

在海量文本数据处理技术中，存储和管理是重要的支撑手段。其中，大容量、可扩展的存储系统是大规模文本数据处理技术的重要保障。目前，大多数文本数据都是采用分布式文件系统（如 HDFS）和分布式数据库（如 Redis）进行存储管理，但随着文本数据规模的不断增大，这些技术已无法满足大规模文本数据处理的需求。因此，在实际应用中需要综合考虑各种因素，构建高效、可扩展、可维护的海量文本数据存储与管理系统。本文介绍了一种基于 Hadoop 的多维存储与管理系统，该系统通过并行化设计提高了系统整体性能，并采用了分布式存储和分布式计算技术[3]。

2 自然语言处理技术

2.1 自然语言处理基础

自然语言处理（Natural Language Processing，NLP）是人工智能中一个重要的分支，是对计算机能够理解、处理的自然语言进行研究的学科。NLP 技术可以分为信息检索、机器翻译、语音识别、文本分类、问答系统等。这些技术可以应用在自动文摘、在线客服、自动摘要和问答系统等方面，也可以在搜索引擎和客户服务方面发挥作用。NLP 技术可以被用于语言的分析、理解和生成，包括文本分类、语义理解、文本生成等。它涉及的主要技术包括：命名实体识别（NER）、词性标注（NET）、词形还原（RDF）、信息抽取（IR）和情感分析等[4]。

2.2 词法分析和句法分析

词法分析就是将句子中的单词按照一定的规则，拆分为词组，再将词组进行匹配的过程。其中，最常用的是“同义词”“词形变化”等规则。句法分析主要包括句子成分分析和句子结构分析，其中，句子成分分析是指将句子拆分为独立的句组。句法结构分析包括：主谓宾、定状补等。其中，主谓宾是句子成分分析中最重要的概念之一。在自然语言处理中，句法和词法分析是两个基本任务，其中句法分析又分为单句（Single-sided）和篇章（Frame）两个层次。词法分析在篇章处理中也占有重要地位。

2.3 信息抽取和文本分类

信息抽取技术是指从文本数据中自动提取信息，它主要包括实体识别、关系抽取和事件抽取三个子任务。信息抽取是一种自然语言处理技术，其任务是从文本中识别出特定的实体、属性或者事件。实体识别是将文本中的单词或短语作为一个整体，根据给定的规则，识别出句子中的名词或代词等实体。关系抽取是从文本中抽取出隐含的、结构化的、语义相关的文本对象。信息抽取技术对大规模文本数据进行分析，从中提取出关键信息，再把这些关键信息转化成计算机能理解和处理的形式[5]。

2.4 深度学习在自然语言处理中的应用

（1）数据预处理阶段：抽取和清洗大量的原始文本数据。（2）特征提取阶段：使用神经网络算法对原始数据进行特征提取，这一过程通常需要大量的训练数据。（3）文本表示阶段：使用神经网络算法对数据进行表示，以更好地进行分类和预测。（4）分类预测阶段：使用神经网络算法对文本数据进行分类预测，实现对文本信息的处理，达到机器理解和判断的目的。深度学习是人工智能领域的一个重要分支，其原理是模拟人类大脑中神经元之间的相互连接以及学习过程，能够对复杂的数据进行处理。

3 情感分析系统设计

3.1 情感分析概述

情感分析是对文本中的观点或情感进行识别、分类和评估，从而为信息用户提供有用的信息服务。情感分析包括基于文本的情感分析和基于图像的情感分析。文本情感分析是将文本作为一种形式，即在给定文本中根据给定的关键词识别出情感极性；图像情感分析是将图像作为一种形式，即从给定的图像中识别出情感极性。通过对用户评论、网页信息等文本数据进行自动标注，将文本信息转化为计算机能够理解和处理的形式，即通过对数据进行标注和训练，从而使计算机能够对数据进行自动识别，进而达到对数据进行自动分类和评估的目的。本系统主要是采用基于规则的方法来实现[6]。

3.2 情感分类算法

情感分类算法在很大程度上决定了情感分析系统的性能，其中包括词向量和特征提取方法等。词向量是指利用词表（text tools）来表示文本，它是对文本进行分词、词性标注的结果。特征提取是指从大量的文本数据中提取出与情感分析任务相关的特征，常用的方法包括 TF-IDF、词袋模型（bag of words）等。词向量和特征提取方法都是对文本进行处理后再进行分类，这两种方法都能使分类准确率得到显著提高，但由于不同算法对样本的要求不同，所以选择适当的算法能使分类准确率达到最佳。

3.3 情感词典构建与情感分析模型

首先，利用互联网上的语料进行情感词典的构建。在此过程中，我们首先从互联网上抓取文本，对文本进行分词、停用词过滤、去除停用词，然后把处理后的文本分为多个部分。通过这些部分构建出情感词典。其次，将这些情感词典作为输入层，基于卷积神经网络、循环神经网络等模型进行情感分类。最后，将训练好的情感分类模型输出的结果作为输出层，利用同样的方法将结果送入下一层，再次进行循环，直到训练完整个模型为止。利用这一方式构建了情感词典与情感分类模型[7]。情感分析的最终目标是获得用户对于产品、服务以及公司等方面的整体评价。

3.4 情感分析系统架构设计

（1）用户端：主要负责用户对情感分析模型的调用，并返回最终的情感分析结果。用户可通过客户端与

服务端进行交互。（2）服务端：主要负责与服务端进行交互，并根据用户返回的结果，对模型进行相应的调整，以获得最优的情感分析效果。（3）服务端与客户端之间通过 HTTP 协议进行交互。客户端向服务端发送请求，服务端在收到请求后，通过 HTTP 协议将处理后的结果返回给客户端。（4）客户端：主要负责返回最终的情感分析结果，并根据用户需要对结果进行展示，同时也可根据用户需求进行修改、删除、导出等操作。

4 实验与结果分析

4.1 数据集介绍

本文在公共情感分析数据集上进行了实验，该数据集包含了来自中国互联网评论语料库（China Online Reviews Content）的评论，该数据集包括了网络上的两种不同的评论语料库，其中一种是针对非中文用户的，另一种是针对中文用户的。除此之外，本文还在此数据集中进行了数据增强，包括对一些缺失值进行填充、删除一些不必要的词等操作，这样处理后的数据集包含了更加丰富的信息。为了评估本文提出的情感分析算法性能，我们从公共情感分析数据集上选取了一些常用的方法作为基准方法，在本实验中我们使用了四种评估指标：准确率、召回率、F1 值和 ROC 曲线。

4.2 实验设计

我们将每个数据集都分为训练集和测试集，其中训练集用于训练模型，测试集用于对模型进行评估。具体来说，我们采用了两种评价指标：F1 值和 Accuracy。F1 值是分类准确度（Accuracy）的度量标准，它是用来衡量模型在所有测试集中分类准确度的平均值。为了比较不同算法的性能，我们采用了基准算法与改进算法的结合。其中，我们在基准算法中采用了无监督学习算法作为对比算法。

4.3 实验结果分析

本文提出的情感分析系统能够较好地对中文文本进行情感分析，并实现了对于文本的情感极性分类，在一定程度上达到了预期效果。而在分类方法的选择上，本文通过基于文本的情感极性分类模型和基于情感词的情感极性分类模型进行对比分析，实验结果显示，基于文本的情感极性分类模型在情感极性分类的准确度上相对较差，而基于情感词的情感极性分类模型则具有较高的准确度。这是因为该系统在进行文本信息挖掘时，直接从用户在社交媒体上所发布的消息中提取出文本信息作为研究对象，因此能够更为准确地对用户所表达的观点进行分析[8]。

4.4 结果讨论

本文的实验结果表明，对于大规模数据集，该系统的表现要比传统的自然语言处理系统有很大的提升，尤其在文本情感分析任务上。与传统的基于规则的情感分析算法相比，本文提出的算法不需要预先设定情感判断规则，也不需要对每个句子进行标注，而是通过训练得到一种自适应的情感分类器，对句子进行自动情感分类。这种方法使得整个系统能够自动识别和处理海量文本数据中的复杂情感分析任务。

5 结语

随着互联网的快速发展，海量文本数据的存储和管理需求日益迫切，而分布式存储与管理系统正是为了解决这些问题而出现的。本文介绍了一种基于 Hadoop 的多维存储与管理系统，该系统通过并行化设计提高了系统整体性能，并采用分布式存储和分布式计算技术，使系统具有良好的可扩展性和可维护性。该系统包括文本数据获取、文本数据预处理、情感分类、情感词典构建和情感分析五个模块。通过实际测试，该系统能够有效地处理大规模文本数据，具有良好的性能表现。随着大数据时代的到来，海量文本数据处理将成为未来计算机研究的一个重要领域。

参考文献

[1]王立栩.全渠道智能客服平台中自然语言处理技术的应用与优化[J].中国宽带,2025,21(09):22-24.
[2]杨海娟。浅析 AIGC 的发展及其在教学中的应用[J/OL]。兰州职业技术学院学报，1-6[2025-08-04].
[3]王晓雨,王灿发.基于自然语言处理（NLP）的网络谣言智能识别与治理框架建构[J].新闻爱好者,2025,(07):21-24.
[4]王超,孙晓宁,孙松.互联网驱动下自然语言处理课程资源库的多维需求分析与实现[J].数字通信世界,2025,(07):247-249.
[5]陈怡.自然语言处理技术下的二语写作语言特征研究：回顾与展望[J].外语教学与研究,2025,57(04):584-595.
[6]石瑞东,邢博,马俊.大语言模型在智慧旅游中的应用——以“智游宁夏”为例[J].信息系统工程,2025,(07):145-148.
[7]孔力,周俊生,张婷.新工科理念下自然语言处理课程教学设计探索[J].计算机教育,2025,(07):214-218.
[8]张驰,徐莉.基于自然语言处理技术的智能客服系统在广电行业的研究应用[J].广播电视网络,2025,32(06):31-33.