

# 基于情感词典与语义规则集的微博文本情感分析

王伟贤, 吴俊

扬州大学信息工程学院, 江苏 扬州

收稿日期: 2023年3月15日; 录用日期: 2023年4月13日; 发布日期: 2023年4月20日

## 摘要

近年来, 以微博为代表的中文媒体平台正在不断融入人们的生活, 人们每天都在这些平台上发表自己的观点、感受等其他主观信息, 如何从这些信息中提取有价值的情感信息并加以利用就称作情感分析。本文提出了一种基于情感词典与语义规则集的微博文本情感分析方法。我们的方法将现有的多个基础情感词典结合起来, 并基于统计信息的方法构建了微博领域情感词典, 同时考虑到中文的语义特性, 加入了自定义的语义规则集。为了验证该方法的有效性, 我们通过网络爬虫技术获取微博中关于新冠肺炎的评论信息共10万条微博文本, 在此数据集上进行了实验。实验结果表明, 与传统的基于情感词典的方法相比, 我们的方法具有更高的准确性和更稳定的表现, 正面、负面和中性情感识别准确率分别达到了79.4%、82.5%、77.3%。综上所述, 本文提出的基于情感词典与语义规则集的微博文本情感分析方法具有较高的准确性和泛化能力, 能够有效地识别微博文本中的情感, 并具有应用价值。

## 关键词

情感分析, 微博文本, 情感词典, 规则集

# Sentiment Analysis of Microblog Text Based on Sentiment Dictionary and Semantic Rule Set

Weixian Wang, Jun Wu

College of Information Engineering, Yangzhou University, Yangzhou Jiangsu

Received: Mar. 15<sup>th</sup>, 2023; accepted: Apr. 13<sup>th</sup>, 2023; published: Apr. 20<sup>th</sup>, 2023

文章引用: 王伟贤, 吴俊. 基于情感词典与语义规则集的微博文本情感分析[J]. 计算机科学与应用, 2023, 13(4): 754-763. DOI: 10.12677/csa.2023.134074

## Abstract

In recent years, Chinese media platforms represented by microblog have been increasingly integrated into people's lives. People express their views, feelings and other subjective information on these platforms every day. How to extract valuable sentimental information from this information and make use of it is called sentiment analysis. In this paper, a sentiment analysis method based on sentiment dictionary and semantic rule set is proposed. Our method combines several existing basic sentiment dictionaries and constructs a microblog domain sentiment dictionary based on statistical information. At the same time, considering the semantic characteristics of Chinese language, we add a custom semantic rule set. In order to verify the effectiveness of this method, we used web crawler technology to obtain a total of 100,000 microblog comments on COVID-19, and conducted experiments on this data set. The experimental results show that compared with the traditional sentiment dictionary-based method, our method has higher accuracy and more stable performance, and the accuracy rate of positive, negative and neutral sentiment recognition reaches 79.4%, 82.5% and 77.3%, respectively. In conclusion, the sentiment analysis method based on sentiment dictionary and semantic rule set proposed in this paper has high accuracy and generalization ability, and can effectively identify the sentiment in microblog, and has application value.

## Keywords

Sentiment Analysis, Microblog Text, Sentiment Dictionary, Rule Set

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

近年来, 随着社交网络的兴起, 微博已经成为人们重要的社交媒体之一。微博的文本信息量大、实时性高、传播速度快等特点, 使其成为了研究者们关注的热点[1]。其中, 微博文本分类是研究的重要方向之一, 目的是将微博文本根据其内容归为不同的类别, 以便更好地进行管理和分析。

微博文本分类中一个关键的问题是如何有效地利用文本中的情感信息。情感是人们对事物的感受、态度和情绪反应, 是人类社交行为中的重要因素。近年来, 情感词典在情感计算中得到了广泛的应用。情感词典是由一些情感词汇构成的词汇表, 每个情感词汇都带有情感极性标签, 表示该词汇所表达的情感是正面的、负面的, 还是中性的。利用情感词典可以快速有效地分析文本中的情感信息。

在国外, 微博情感分析研究已经开始相当早, 并且得到了很大的发展。其中, 研究者使用的算法和模型也越来越先进。2013年, Mohammad 等人[2]使用情感词典和机器学习算法对英文推文的情感进行分类, 取得了较好的结果。2014年, Mohammad 等人[3]提出了 SentiStrength, 这是一种常用的英语情感分析工具。该算法可以分析文本中的正面和负面情感, 并对其打分。此后, 一些研究者也开始使用类似的方法对微博情感进行分析。在2018年, 一些研究者通过比较不同的算法, 得出了在微博情感分析中表现最好的是基于循环神经网络(RNN)的算法[4]。而在2019年, 研究者们利用预训练的语言模型, 进一步提高了微博情感分析的效果。这些预训练的语言模型主要有 BERT 和 XLNet 等[5]。

在国内, 微博情感分析研究也在逐渐发展。2003年, 朱嫣岚等人[6]提出了基于 HowNet 词汇语义相似度和语义相关场的情感词极性计算方法, 文本判别准确率可达 80%以上。2009年, 王素格等人[7]提出

了基于同义词的词汇情感倾向判别方法,将目标词语与其同义词的关系考虑到其中,取得了不错的效果。在 2016 年,刘龙飞等人[8]提出了一种基于卷积神经网络(CNN)的微博情感分析模型。该模型通过学习微博中的局部特征,可以对微博进行情感分析,并且取得了不错的效果。在 2017 年,姜杰等人[9]提出一种机器学习与规则相融合的微博情感分类方法,将用规则方法得到的多样化情感信息进行转化,扩展并嵌入基本特征模板,形成更有效的融合特征模板。实验结果表明,该方法比传统的基于机器学习的方法更加有效。在 2018 年,凌海彬等人[10]提出了一种基于多特征融合的微博情感分析方法。该方法将微博中的文本内容特征特征和用户特征进行融合,并且设计了特征层融合和决策层融合方法进行图文融合的情感分类,进一步提高了微博情感分析的准确率。

微博情感分析通常是采用情感词典、机器学习和深度学习这三类方法[11],它们都具有一定的有效性。然而,无论哪种方法都无法达到最高的准确率,只能通过不断改进方法来提高准确性。其中,基于情感词典的方法适用于处理细粒度的文本情感分析,因此本文主要采用情感词典方法并结合语义规则集来对微博进行情感分析。在此基础上,不断改进和优化,通过各个部分的情感权值加权求和来得到微博的情感极性[12]。

本文的结构如下:第 2 节介绍了本研究实验所使用研究方法,包括使用的数据集、情感词典的构建以及语义规则集的定义。第 3 节介绍实验的流程、实验数据结果。第 4 节总结了论文,包括所使用方法的优劣势以及后续工作的讨论。

## 2. 方法

### 2.1. 数据集

本次实验所使用的数据集主要来源于微博评论中与新冠肺炎相关的文本。我们使用网络爬虫技术,从微博中获取了共计 10 万条关于新冠肺炎的评论信息(如图 1 所示)。为了获取微博评论信息,采用了特定的程序来捕捉和提取微博评论中的信息,包括微博 id、发布者昵称、微博正文、发布时间、点赞数、转发数、评论数等,并将获取到的数据保存在本地 CSV 格式文件中,形成了本文所使用的数据集“Virus\_W”。为了进行情感标注分类,使用了人工手段进行标注。由于标注结果因人而异,因此邀请多位标注员对同一数据集进行标注。当标注结果不一致时,按照少数服从多数的原则,选取具有压倒性优势的标注作为最终标注结果。

回复@李明的一说春秋:有意思,那我还真就要计较下,我有一句话	-1
#全国确诊新型肺炎病例#加油 武汉加油 中国[心]??	1
多注意	1
1.22[心]	1
【#上海确诊1例新型冠状病毒感染肺炎# 自武汉来沪】1月20日晚,	0
[鼓掌]	1
//@斯奇就是我://@汪苏泷·光影星云:#武汉加油# 我们一起,都要好	1
//@午后狂睡:我们不能只有在发生疫情的时候才想起来要珍惜医护人	-1
我妈看新闻的时候嚷嚷不出门不出门,结果明天还是要带我回外婆家	-1
这几位是孤儿吧,医院职工大年三十忙了一天,晚饭一桶干瘪泡面,	-1
崽崽别担心,武汉一定能挺过去的! //@朱一龙:#武汉加油#我们在	1
真好, happy everyday	1
//@协和医生Do先生:这是最新医院名单,恳请转发出去. 今天是大年	-1

Figure 1. Experimental data sample

图 1. 实验数据示例

### 2.2. 数据预处理

考虑到微博评论数据的不规范性,含有大量噪声,需要做数据预处理。下图 2 所示是数据预处理流

程。

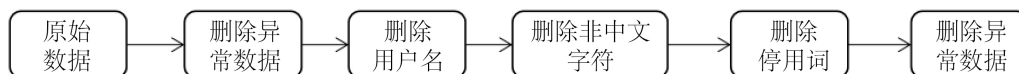


Figure 2. Data preprocessing process

图 2. 数据预处理流程

### 2.3. 构建基础情感词典

本文采用基于情感词典的情感分析方法, 关键是构建一个完整的情感词典, 以便能够满足情感分析的需求。鉴于现代网络社会中新词不断涌现, 情感词典需要进行扩充。本文主要将知网的 HowNet 词典、大连理工大学整理的权威情感词汇本体库(DUTSO)和台湾大学整理的 NTUSD 词典结合起来构建基础情感词典, 合并去重后共计 43783 个情感词汇。同时, 考虑到中文的语义特性, 还加入了否定词词典和程度副词词典, 如表 1、表 2 所示。

Table 1. Examples of negative words

表 1. 否定词示例

词语类型	词语	权值
否定词	不、无、否、没……	-1
双重否定词	并非、不是不、并不是不……	1
反问词	难道、莫非、奈何、怎样……	-2

Table 2. Examples of adverb of degree

表 2. 程度副词示例

级别	词语	权值
极 extrem/最 most	极度、极端、绝对、完全	4
很 very	多么、分外、格外、很是	3
更 more/较 more	更加、较为、愈发、越是	2
稍 ish	稍微、稍稍、略微、一点	1

### 2.4. 构建微博新词情感词典

我们将微博博文作为语料, 对话料数据进行数据清理和数据降噪, 然后对话料数据进行分词、去除停用词和删除其中的非情感词语。然后通过基于统计信息的方法识别出新词, 对得到的词语与基础情感词典中的词进行对比去重, 得到候选词, 构成候选情感词典。将候选词按照词频排序, 取情感极性强度和词频高的作为种子词(正负各 30 个), 人工判断这些种子词的词性。利用 SO-PMI 算法判断候选情感词的情感极性, 完成基础情感词典的扩充。图 3 为微博新词的情感词典的构建流程。

基于统计信息的方法主要是对字串频率、字串内部紧密度以及字串外部邻字信息熵进行标准设定, 这三个定义关系到了一个字串能否成词, 如果这三个定义都达到设定的阈值, 就可以认为该字串成为一个新词, 但是还要比对是否这些词已经存在已有的初始情感词典中, 如果存在就删除, 如果不存在, 就可以认为是一个新词。

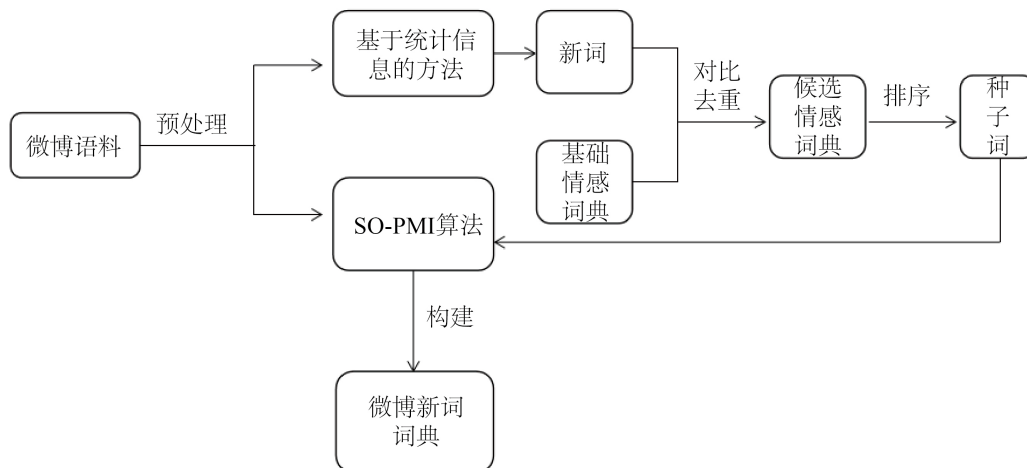


Figure 3. Flow chart of construction of microblog neologism dictionary  
图 3. 微博新词词典构建流程图

SO-PMI 算法由两部分组成: SO-PMI 和 PMI。算法整体思想很简单, 判断需要判断的词语 P (word) 与基准词 P (base)同时出现的概率, 如果与积极(positive)的词同时出现的概率更高, 那么就判断为积极的词语, 如果与消极(negative)的词同时出现的概率更高, 那么就判断为消极的词语, 如果与积极和消极的概率相同, 那么就判断为中性的词语。

对于词语来说, 点间互信息(PMI)可以通过判断某个词与基准词出现的概率来计算词语间的语义相似度。两个词语  $w_1$  和  $w_2$  点间互信息计算公式为:

$$PMI(w_1, w_2) = \log \frac{P(w_1, w_2)}{P(w_1) \times P(w_2)} \quad (1)$$

式中,  $P(w_1, w_2)$  表示  $w_1$  和  $w_2$  同时出现在语料中的概率,  $p(w_1)$  和  $p(w_2)$  分别表示  $w_1$  和  $w_2$  单独出现的概率。如果两者独立, 那么  $P(w_1, w_2) = P(w_1) \times P(w_2)$ , 即整个分数为 1, 那么  $PMI = 0$ 。

将 PMI 算法应用到未知情感极性的新词的情感判断上。用  $w_1$  表示未知情感极性的新词,  $w_2$  表示已知情感极性的词[12], 通过计算两者之间的语义相似度, 如果计算结果很高, 则表示两者之间情感极性一样, 如果语义相似度不高, 则表示两者情感极性相反。

SO-PMI 算法可以进行一个未知新词与多词进行相似度计算, 具体定义如下:

构建正面情感种子词集合  $w_p$  和负面情感种子词集合  $w_n$ , 对每一个新词计算与这些集合中种子词的语义相似度, 从而判断未知新词的情感极性[12], 由以下公式计算:

$$Sen\_PMI_{(w)} = \sum_{w_n \in W_p} PMI(w, w_p) - \sum_{w_n \in W_n} PMI(w, w_n) \quad (2)$$

式中, 左边表示新词  $w$  的情感值, 公式右边表示的是新词  $w$  与正面种子词集合  $w_p$  中每个词的语义相似度之和然后减去新词  $w$  与负面种子词集合  $w_n$  中每个的语义相似度之和。

SO-PMI 的基本思想是: 有两个情感倾向的列表, 一个是正向的情感词, 一个是负向的情感倾向。计算每个词与这两个列表中的每个词的 PMI, 看这个词更接近哪个列表, 是更接近正向情感词列表还是负向情感词列表。若  $SO-PMI > 0$ , 表示这个词更接近正向, 即积极词; 若  $SO-PMI = 0$ , 表示这个词为中性词; 若  $SO-PMI < 0$ , 表示这个词更像是负面词, 即消极词[13]。

对新词进行情感极性判别后将判别出带有情感色彩的词加入基础情感词典中, 完成微博新词情感词典的构建。

## 2.5. 文本情感极性判断规则

### 2.5.1. 词组合规则

情感词、否定词和程度副词的不同组合将会产生不同的情感表达效果。因此, 本章根据不同的词组合模式, 制定相应的情感极性判断规则。

规则 1: 仅情感词。如果只有情感词, 没有其他词修饰情感词, 则该词的情感极性为它本身的情感极性。

规则 2: 程度副词 + 情感词。情感极性不变, 情感强度的变化取决于程度副词。

规则 3: 否定词 + 情感词。在这种词组合中, 情感极性取决于否定词的个数, 情感强度不变。当否定词的个数为偶数时, 情感极性不变; 当否定词的个数为奇数时, 情感极性变为相反极性[14]。

规则 4: 否定词 + 程度副词 + 情感词。在这种词组合中, 否定词修饰程度副词, 程度副词修饰情感词, 因此其情感极性不变, 情感强度的变化取决于否定词的个数。

规则 5: 程度副词 + 否定词 + 情感词。这种词组合中, 否定词修饰情感词, 因此其情感极性是否改变取决于否定词的个数, 情感强度的变化取决于程度副词。

### 2.5.2. 句型规则

中文按句型可以大致分为陈述句、疑问句、反问句和感叹句[15], 如图 4、图 5 所示, 不同的句型对应的情感强度及情感极性各不相同。于是, 我们对不同句型制定相应的规则并赋予权重。

须警惕, 不轻视! 认真做好防护工作, 我们一起加油!

**Figure 4.** Examples of declarative and exclamatory sentences

**图 4.** 陈述句、感叹句举例

每当看到疫情消息, 不知道为什么会感觉鼻子一酸, 又心疼又难过, 在国外念书, 看到自己的祖国自己的家里发生这些事情, 想做些什么, 又做不到些什么, 我是个男孩子, 难道这就是软弱无力吗?

**Figure 5.** Examples of rhetorical question

**图 5.** 反问句举例

一个完整的句子一般都是复句, 因此除了从复句内部的各个单句考虑之外, 还要从整个复句的角度去考虑, 也就是复句的句型。复句的句型直观上的标志是以标点符号结尾的句子, 通过标点符号来判断复句的句型, 比如以“?”结尾的一般事疑问句或者反问句, 以“。”结尾的是陈述句, 以“!”结尾的是感叹句等等。复句与复句的组合构成了一个文本, 本节研究的是句型关系对微博文本的情感影响, 在本节中用字母  $D$  表示一个文本, 文本中各个复句的集合定义为  $\{D_1, \dots, D_i, \dots, D_n\}$  用字母  $T_i$  表示句型规则对复句  $D_i$  的情感值。

规则 1: 感叹句型。文本中, 若复句以“!”结尾, 则该句就是一条感叹句, 此时情感值定义为 1.5。它所表达的情感是将原来的程度加强。

规则 2: 反问句型。文本中, 反问句型存在两种情况。第一种是反问句型以“?”结尾且结尾处出现反问标志词; 第二种是反问句型没有以“?”结尾但结尾处出现反问标志词。此时情感值定义为-1。它所表达的情感是与原来极性相反。

规则 3: 疑问句型。文本中, 疑问反问句型以“?”结尾但结尾处没有出现反问标志词, 比如“难道不应该喜欢我?”, 该句中存在“?”但没有出现“吗”这类反问标志词。因此, 由疑问句型对复句  $D_i$  产

生的情感值定义为 0。它所表达的情感是极性变为中性。

规则 4: 陈述句型。文本中, 若复句以“。”结尾, 则该句就是一条陈述句, 此时情感值定义为 1。它所表达的情感是不变。

### 2.5.3. 句间规则

复句主要由两个及以上的分句通过关联词(如表 3 所示)连接构成。不同关联词引导的分句侧重点和情感均有不同, 关联词示例见表 3。本文根据关联词制定相应规则并赋予权重。用字母 A 表示一个复句, 复句中各个分句的集合定义为  $\{B_1, \dots, B_i, \dots, B_n\}$  用字母  $S_i$  表示句型规则对分句  $B_i$  的情感值。

Table 3. Examples of related words

表 3. 关联词示例

关联词类型	示例
表转折关系	但是, 然而
表递进关系	不但……而且, 并且
表假设关系	如果, 要是
表因果关系	因为……所以, 是因为

规则 1: 转折关联词。转折关联词引导的分句, 前后意思完全相反。因此若转折关联词为后接词(如“但是”“可是”等), 则其引导的后分句  $B_i$  及之后的分句权重  $S_i$  均赋为 1, 前分句权重  $S_i$  均为 0; 若转折关联词为前接词(如“虽然”, “尽管”等), 则其引导的后分句  $B_i$  及之后的分句权重  $S_i$  均赋为 0, 前分句权重  $S_i$  均为 1; 若既有转折前接词也有转折后接词, 则后接词出现的分句  $B_i$  及之后的分句权重均赋为 1,  $B_i$  之前的分句权重均赋为 0。

规则 2: 递进关联词。递进关联词引导的分句, 后分句的情感强度比前分句强。因此分句的权重赋值为:  $S_i = 1, S_{i+1} = 1.5, S_j = 1 + 0.5 \times (j - i)$ 。

规则 3: 假设关联词。假设关联词引导的分句, 其情感极性倾向于前提条件, 削弱了后面分句的语气。因此将假设关联词出现的分句及之前的分句权重均赋值为 1, 之后的分句均赋值为 0.5。

规则 4: 因果关联词。因果关联词引导的分句, 前分句和后分句分别解释了原因和结果, 句子的关键信息在结果分句上。因此将因果关联词中的结果关联词(如“所以”)所在的分句及之后的分句权重均赋值为 1, 之前的分句均赋值为 0.5。

### 2.5.4. 综合情感极性判断

在判断微博文本的情感极性时, 首先根据复句中的标点符号把文本切分成多个分句, 然后计算分句的情感分数, 接着计算整个复句的情感分数, 最后根据该分数判断该段文本的情感极性。

#### 1) 计算分句的情感分数

根据扩充的情感词典匹配情感词, 若分句中没有情感词则该分句情感分数为 0; 若分句中存在情感词, 则根据上文提出的词组合规则, 计算词组合的情感分数, 该分句的情感分数与该词组合情感分数一致。

#### 2) 计算复句的情感分数

在得到分句的情感分数后, 将分句情感分数求和, 并根据上文提出的句型规则, 计算整个复句的情感分数。将复句的情感分数求和得到整段微博文本的情感分数。

#### 3) 判断微博文本的情感极性

若情感分数大于 0, 则微博文本的情感极性为正面; 若情感分数等于 0, 则微博文本的情感极性为中

性; 若情感分数小于 0, 则微博文本的情感极性为负面。

### 3. 实验

#### 3.1. 实验设置

本文主要通过三组实验对比验证提出方法的有效性, 第一组实验是在基础情感词典(知网的 HowNet 词典)上对微博文本数据进行情感分析, 第二组实验是在本文所构造的多维情感词典上对微博文本数据进行情感分析, 第三组实验是在本文所提出的基于情感词典与自定义规则的方法上对微博文本数据进行情感分析。

对于二分类问题, 可将样本根据其真实类别与分类器预测类别的组合划分为真正例(TP)、假正例(FP)、真反例(TN)和假反例(FN)四种情形[16]。在本实验中, 我们主要使用以下三个指标进行分析, 即准确率 P、召回率 R 和综合指标 F1。它们通过以下三个公式进行计算。

1) 准确率(Precision)亦称“查准率”, 为预测为正例且正确预测的样本数与所有预测为正例的样本数的比值[16]:

$$P = \frac{TP}{TP + FP} \quad (3)$$

2) 召回率(Recall)亦称“查全率”, 为预测为正例且正确预测的样本数与真实情况下所有为正例的样本数的比值[16]:

$$R = \frac{TP}{TP + FN} \quad (4)$$

3) F1 值是基于查准率与查全率的调和平均定义的[16]:

$$F1 = \frac{2 * P * R}{P + R} \quad (5)$$

#### 3.2. 实验结果与分析

通过对数据集分别进行三组实验, 并比较了实验结果, 三个评价指标结果如下表 4、表 5 和表 6 所示。

**Table 4.** Accuracy rate of sentiment analysis of microblog text data under three different methods

**表 4.** 三种不同方法下对微博文本数据进行情感分析的准确率

	基础情感词典	多维情感词典	多维情感词典与自定义规则
正面	60.10%	73.10%	79.40%
负面	62.40%	75.80%	82.50%
中性	58.30%	71.20%	77.30%

**Table 5.** Recall rate of sentiment analysis of microblog text data under three different methods

**表 5.** 三种不同方法下对微博文本数据进行情感分析的召回率

	基础情感词典	多维情感词典	多维情感词典与自定义规则
正面	58.70%	69.20%	77.40%
负面	63.10%	76.70%	83.20%
中性	65.60%	78.40%	81.60%



**Table 6.** F1 value of sentiment analysis of microblog text data under three different methods**表 6.** 三种不同方法下对微博文本数据进行情感分析的 F1 值

	基础情感词典	多维情感词典	多维情感词典与自定义规则
正面	60.70%	73.10%	80.20%
负面	62.30%	75.60%	83.70%
中性	59.20%	71.70%	75.40%

通过以上的结果呈现可以看出, 本文提出的基于多维情感词典与自定义规则的微博文本情感分析方法具有一定的效果, 相较于前面两种方法都有很好的提升。在只有一部基础情感词典的基础上, 远远达不到中文微博文本的情感词匹配要求, 正是由于本文构造了微博新词情感词典, 再加上其他几部词典才能有很好的作用。此外, 还需要将中文文本语义规则集考虑其中, 只有这样, 整个微博的情感分析才能进一步提升, 因为文本是表示含义的, 是有语法结构的, 因此需要结合规则一起进行情感分析, 上表中从准确率、召回率和 F 值三个层面分析可知本文提出的方法相比于前两种方法有显著的效果。

#### 4. 总结

在本研究中, 我们提出了一种新的基于微博文本情感分析的方法, 用于识别文本的正面、负面和中性情感特性。实验使用从微博爬取的关于新冠疫情的数据集进行测试, 并将以往的研究结果与本研究提出的方法进行比较。此方法的重点在于, 在原始情感的基础上, 加入了构建的微博新词情感词典, 并定义了适合中文文本的语义规则集, 提高了文本情感识别的准确率。

表 4、表 5、表 6 中的情感识别研究均基于中文微博文本, 通过对比发现, 目前已有的情感词典无法满足微博情感分析, 使用多维情感词典的方法普遍优于使用基础情感词典的方法。同时, 加入结合微博文本的语言特性定义的规则集也可以有效地提升最终结果的准确率。通过表中准确率的数据对比, 本文的研究方法在正面、负面和中性文本的情感识别准确率分别达到了 79.4%、82.5%、77.3%。

当前工作的局限性及相应的未来研究方向总结如下:

- 需要人工判断新增情感词语的强度, 工作量太大。
- 未考虑表情符号在实际评论中隐藏的情感属性, 后续可考虑将表情符号转换为相应的文字, 加入到情感词典中。
- 对于不同的事件而言, 文本中的情感词语表达的含义可能不同, 应考虑词语的多重情感特性。

本文的情感分析仅针对微博文本, 实际上各类网络平台上的文本包含了大量情感信息, 将本文提出的情感分析方法应用到其他数据集上将是下一步的研究重点。

#### 参考文献

- [1] 高雅, 苏艳, 席方园. 基于 Python 的新浪微博用户数据采集与分析[J]. 电子设计工程, 2019, 27(20): 157-160+165.
- [2] Mohammad, S.M., Kiritchenko, S. and Zhu, X. (2013) NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. ArXiv: 1308.6242.
- [3] Mohammad, S.M. and Turney, P.D. (2013) Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29, 436-465. <https://doi.org/10.1111/j.1467-8640.2012.00460.x>
- [4] Devlin, J., Chang, M.W., Lee, K., et al. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. ArXiv: 1810.04805.
- [5] Yang, Z., Dai, Z., Yang, Y., et al. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Vancouver, 8-14 December 2019, 5753-5763.

- 
- [6] 朱嫣岚, 闵锦, 周雅倩, 等. 基于 HowNet 的词汇语义倾向计算[J]. 中文信息学报, 2006, 20(1): 14-20.
- [7] 王素格, 李德玉, 魏英杰, 等. 基于同义词的词汇情感倾向判别方法[J]. 中文信息学报, 2009, 23(5): 68-74.
- [8] 刘龙飞, 杨亮, 张绍武, 等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报, 2015, 29(6): 159-165.
- [9] 姜杰, 夏睿. 机器学习与语义规则融合的微博情感分类方法[J]. 北京大学学报(自然科学版), 2017, 53(2): 247-254.
- [10] 凌海彬. 基于多特征融合的微博情感分析研究[D]: [硕士学位论文]. 桂林: 桂林电子科技大学, 2019.
- [11] 王婷, 杨文忠. 文本情感分析方法研究综述[J]. 计算机工程与应用, 2021, 57(12): 11-24.
- [12] 吴杰胜, 陆奎. 基于多部情感词典和规则集的中文微博情感分析研究[J]. 计算机应用与软件, 2019, 36(9): 93-99.
- [13] 谈光璞, 朱广丽, 韦斯羽. 基于情感特征增强的中文隐式情感分类模型[EB/OL]. 计算机工程与应用: 1-13. <http://kns.cnki.net/kcms/detail/11.2127.TP.20221128.0932.010.html>, 2022-11-18.
- [14] 曹东伟, 李邵梅, 陈鸿昶, 等. 融合情感特征的虚假评论检测方法[J]. 信息工程大学学报, 2021, 22(3): 326-330.
- [15] 徐康庭, 宋威. 结合语言知识和深度学习的中文文本情感分析方法[J]. 大数据, 2022, 8(3): 115-127.
- [16] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.