微博用户情感分析

摘要

随着互联网的普及，互联网中公开可用的信息不断增长，在论坛、博客等社交媒体中可以获得大量用户表达观点的文本数据，这些数据可以帮助刻画用户画像。本文以新浪微博为例，通过爬取指定用户的全部微博进行词频分析、词云图绘制以及基于情感词典的情感分析，判断微博中每条观点句的情感倾向，进行微博用户情感分析。

1 项目需求

本项目从微博用户出发，可分为以下几个步骤，其对应的需求如下：

①数据采集：编写定向爬虫程序，爬取用户的所有微博并存放在.csv 文件中。

②预处理：使用 jieba 对进行微博文本进行分词，导入停用词对分词结果进行筛选进行预处理。

③数据分析：根据分词结果统计词频，导入情感词典、否定词词典等进行情感分析，计算每条微博的情感得分，按照是否原创、情感为积极或消极计算情感均值和方差。

④结果可视化：根据词频结果绘制词云图，情感得分绘制直方图，实现结果可视化。

⑤评估检验：人工对微博对应的情感值进行积极与消极标记，与得到的情感得分相比较。

2 难点问题及相关方法

①微博爬虫：微博的反爬机制比较强，爬取微博桌面端、微博触屏版这种动态加载的网页不能一次抓取到所有内容，而重复爬取会造成微博账号被封以至于不能继续爬取网页。

爬取 url 换为 https://weibo.cn/，这是一个比较老的网页，预先登录已注册账号获取 cookies，为避免多次重复爬取，将 html 源码存储到文件中，设置爬取间隔避免操作过于频繁。

②停用词处理：由于微博文本的特殊性，用户发布的微博中包含以“@”(提醒某人)标识用户名信息，以“##”标识的话题信息，以“***的微博视频”为后缀的视频、组图等等，jieba 分词不能对特定名词进行正确划分，比如“易烊千玺”，jieba 分词结果为[‘易’, ‘烊’, ‘千玺’]，而这些文本对用户来说又没有实际意义，影响词频统计和情感分析。

由于不同用户的偏好不同，对所有用户构建相同的停用词列表任务繁重，而同一用户的偏好基本不会发生很大变化。所以在本项目中，针对每一用户，先导入停用词列表对 jieba 分词进行初步筛选，词频统计，在得到的词频文件中，人工筛选出适合该用户的停用词列表，进行二次筛选。

③emoji 的处理：在微博文本中，有许多 emoji 表情，这些 emoji 表情同样表达了用户的情感。而在现有的情感词典中很少包含 emoji，因此，如何对 emoji 进行处理成了难点之一。

汉语词典 emoji 对照表[6]中有大量的 emoji 及其对应的汉语含义，可以爬取相关数据构建 emoji 词典，在 jieba 分词后对分词结果中的 emoji 进行中文替换再进行导入停用词等后续步骤，以完成 emoji 的处理。

3 关键技术或主要模型实现

①数据采集：使用 requests 方法对 html 页面进行爬取，使用 xpath 和正则表达式对微博数据进行提取，由于原创微博、转发微博、长微博等的 html 标签不同，所以需要编写不同的解析函数对微博进行解析，最终将爬取到的数据(微博 id、微博正文/转发理由、是否原创)存放在.csv 文件中供后续使用。

②预处理：将同一的所有微博用空格连接，使用 jieba 对其进行分词，导入停用词和 emoji 词典，遍历所有分词结果中的所有词语，进行初步筛选和 emoji 替换，筛选后统计词频。根据词频结果更新停用词列表，再次进行词频统计。

③数据分析：对每一条微博进行 jieba 分词得到分词列表，并导入②中得到的停用词列表进行筛选。导入情感词典、程度副词词典和否定词词典，在筛选后的分词列表中的词语按照是否在情感词典等词典中进行分类，找出列表中的情感词、程度副词和否定词，将三种类型的词语分别存放在三个 Python 字典中，key 的值为单词在分词列表中的下标，value 的值为单词在词典中对应的值，否定词的值为-1。得到分类后的词典后，进行基于情感词典的微博文本情感分析，得到每条微博的情感得分，并把得分写入到文件中。

基于情感词典的微博文本情感分析方法如下：初始情感得分为 0，初始权重为 1，遍历情感词字典，查看两个情感词之间是否有程度副词和否定词，若有程度副词，则权重为权重与程度数值乘积，若有否定词，则权重为权重*(-1)^否定词个数，情感得分为权重与情感值乘积的累加和。

④结果可视化：根据词频统计绘制词云图，根据情感得分绘制直方图，实现结果可视化。

⑤评估检验：人工对微博对应的情感值进行积极与消极标记，将真实情感与分析出的情感进行比较，计算准确率。

4 实验结果及分析

4.1 评价标准

定义情感均值对微博用户的积极消极向进行评价，情感均值的定义为：

$$Mean = \frac{\sum_{i=1}^{n} Score}{n}$$

其中，*n* 表示数据集中微博用户所有微博总数，*Score* 表示每条微博情感得分。

定义情感方差对情感得分离散程度进行评价，情感方差的定义为：

$$\text{Variance} = \frac{\sum_{i=1}^{n}\left\{\sum_{j=1}^{n}\{(Score - Mean)^2\}\right\}}{n}$$

其中，*n* 表示数据集中微博用户所有微博总数，*Score* 表示每条微博情感得分，*Mean* 表示情感均值。

定义准确率对实验结果的正确率进行评价，准确率的定义为：

$$Accuracy = \frac{N_t}{N_r}$$

其中，*Nt* 表示被正确判断的样本数，*Nr* 表示总样本数。

4.2 数据集

本文中用到的用户微博数据集为数据采集部分爬取的微博数据[1] [2]，本文中用到的情感词典为 BosonNLP 情感词典[3]，停用词典为 SnowNlp 提供的停用词词典[5]，否定词词典为科学空间提供的否定词词典[6]，程度副词词典为《知网》情感分析用词语集[4]，emoji 词典为对汉语词典 emoji 对照表[7]的爬取结果。下面从数据集大小和数据特点等方面介绍 2 个微博数据集和 4 个词典。

表 1 为本次实验所用用户微博数据集信息：

表 1 用户微博数据集

数据集	微博总条数	原创条数	转发条数
TFBOYS-易烊千玺	828	516	312
老番茄	71	64	7

人工对用户每条微博情感值进行标记，得到带有情感标记的数据集信息如表 2：

表 2 带有情感标记的用户微博数据集

数据集	微博总条数	积极条数	消极条数
TFBOYS-易烊千玺	828	811	17
老番茄	71	39	32

用户微博数据集的内容为用户名为“TFBOYS-易烊千玺”[1]和“_老番茄_”[2]的全部微博，包含微博 id、微博正文/转发理由等信息，格式为[[微博 id][微博正文/转发理由][是否原创]]，每行表示一条微博的相关信息，以逗号分隔，存储在.csv 文件中。其中，用户名为“TFBOYS-易烊千玺”的微博数据集中共有 828 条微博，512 条原创微博，316 条转发微博；用户名为“_老番茄_”的微博数据集共有 71 条微博，64 条原创微博，7 条转发微博。

表 3 为本次实验所用有关词典信息：

表 3 词典信息

词典名称	词语总条数	最低情感分值	最高情感分值
BosonNLP 情感词典	114766	-6.7040	6.3750
程度副词词典	212	0.5	1.8
否定词词典	71	-1	-1
停用词词典	1910	--	--
emoji 词典	765	--	--

BosonNLP 情感词典[3]是从微博、新闻、论坛等数据来源的上百万篇情感标注数据当中自动构建的情感极性词典，格式为[[词语]] [[情感值]]，每行为一个情感词及其对应的情感分值，以空格分隔，共包括 114767 个词语。其中负数代表偏负面的词语，非负数代表偏正面的词语，正负的程度由数值的大小反应出。

程度副词词典词典[4]格式为[[程度副词] [程度数值]]，每行为一个程度副词及其对应的程度数值，以空格分隔，共包括 212 个词语，其中> 1 表示强化情感，< 1 表示弱化情感。

停用词词典[5]共包含 1893 个停用词，每行表示一个停用词。

否定词词典[6]共包含 71 个否定词语，每行表示一个否定词。

emoji 词典[7] 为对汉语词典 emoji 对照表[7]的爬取结果，格式为[[emoji 表情] [汉语含义]]，每行表示一个 emoji 及其汉语含义，以空格分隔，共包含 765 个 emoji 表情。

4.3 实验结果

在本节中，进行实验对词频进行分析绘图，对用户的每条进行情感值打分，计算用户的平均情感值与方差等，并于人工标注的微博情感对比，计算准确率。

通过进行词频分析，绘制两个微博用户的词云图如图 1 和图 2：

图 1 微博用户“TFBOYS-易烊千玺”词云图

图 2 微博用户“_老番茄_”词云图

通过绘制两名用户的词云图可以分析出，微博用户“TFBOYS-易烊千玺”微博中经常出现的词语有“TFBOYS”、“开心”、“谢谢”、“支持”等，可以推测出其作为公众人物的身份，微博用户“_老番茄_”微博中经常出现的词语有“百大”、“哔哩”、“明天”、“作业”、“头发”等，可以推测出其作为哔哩哔哩 up 主及大学生的身份。

对两名用户的所有微博进行情感打分，打分结果统计如表 4-1、表 4-2、表 4-3、表 4-4、表 4-5，用 matplotlib 模块绘制频数统计直方图，微博用户“TFBOYS-易烊千玺”的统计结果如图 3-图 7，微博用户“_老番茄_”的统计结果如图 7-图 12：

表 4-1 基于情感词典的微博用户情感得分分析（频数分析）

数据集	微博总条数	积极条数	消极条数	积极频率	消极频率
TFBOYS-易烊千玺	828	723	72	0.8732	0.0870
老番茄	71	39	29	0.5493	0.4085

表 4-2 基于情感词典的微博用户情感得分分析（总体）

数据集	微博总条数	最低情感得分	最高情感得分	情感均值	情感方差
TFBOYS-易烊千玺	828	-21.5388	56.1020	8.3861	81.3360
老番茄	71	-13.5270	30.6026	1.5000	39.1442

表 4-3 基于情感词典的微博用户情感得分分析（积极）

数据集	积极条数	最低情感得分	最高情感得分	情感均值	情感方差
TFBOYS-易烊千玺	723	0.0126	56.1020	9.8108	75.931
老番茄	39	0.1470	30.6026	4.7870	40.8298

表 4-4 基于情感词典的微博用户情感得分分析（消极）

数据集	消极条数	最低情感得分	最高情感得分	情感均值	情感方差
TFBOYS-易烊千玺	72	-21.5388	-0.0360	-2.0074	10.7798
老番茄	29	-13.5270	-0.1938	-2.7651	7.9730

表 4-4 基于情感词典的微博用户情感得分分析（原创）

数据集	原创条数	最低情感得分	最高情感得分	情感均值	情感方差
TFBOYS-易烊千玺	516	-21.5388	56.1020	8.7066	90.5244
老番茄	64	-13.5270	14.9158	0.2955	16.6694

表 4-5 基于情感词典的微博用户情感得分分析(转发)

数据集	转发条数	最低情感得分	最高情感得分	转发均值	转发方差
TFBOYS-易烊千玺	312	-4.8317	52.8646	7.8559	65.6889
老番茄	9	0.0000	30.6026	12.5136	109.0587

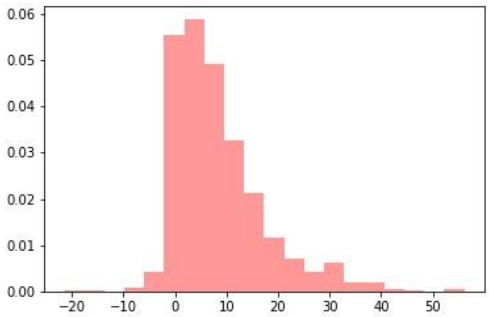

图 3 微博用户“TFBOYS-易烊千玺”全部微博情感值频数分布直方图

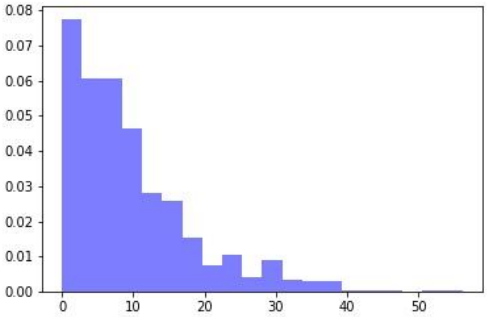

图 4 微博用户“TFBOYS-易烊千玺”积极微博情感值频数分布直方图

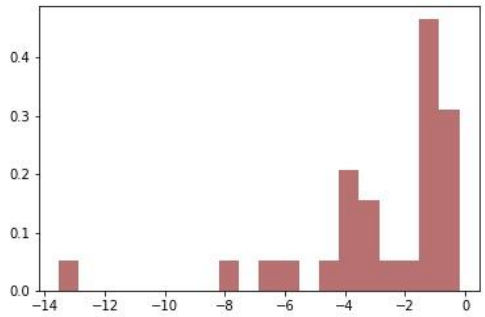

图 5 微博用户“TFBOYS-易烊千玺”消极微博情感值频数分布直方图

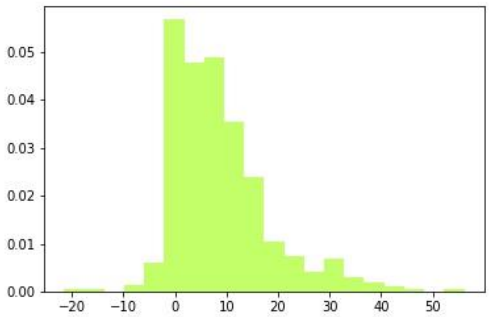

图 6 微博用户“TFBOYS-易烊千玺”转发微博情感值频数分布直方图

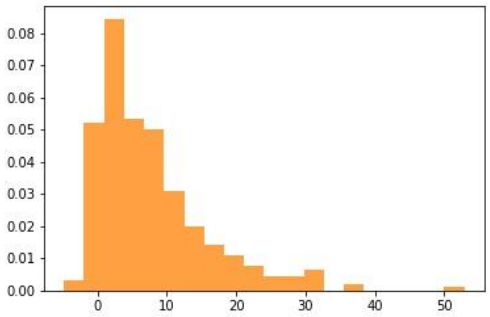

图 7 微博用户“TFBOYS-易烊千玺”转发微博情感值频数分布直方图

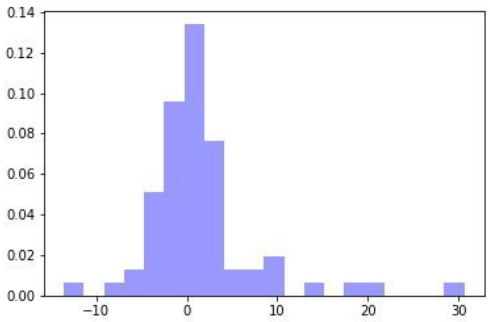

图 8 微博用户“_老番茄_”全部微博情感值频数分布直方图

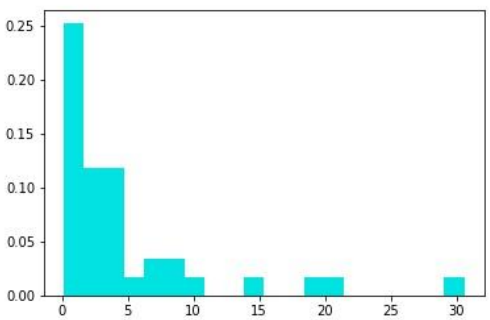

图 9 微博用户“_老番茄_”积极微博情感值频数分布直方图

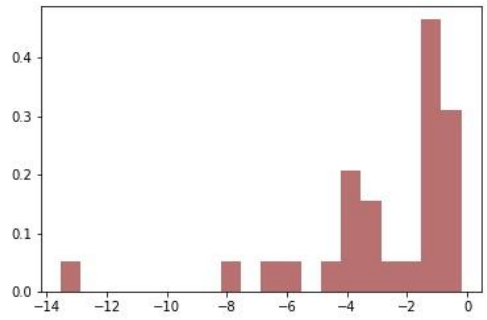

图 10 微博用户“_老番茄_”消极微博情感值频数分布直方图

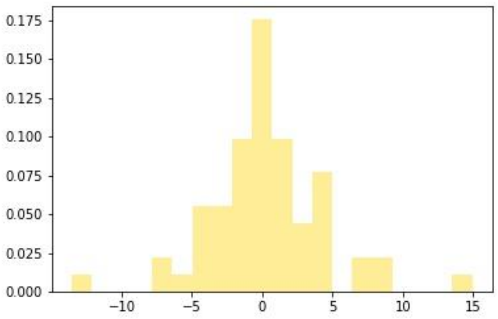

图 11 微博用户“_老番茄_”原创微博情感值频数分布直方图

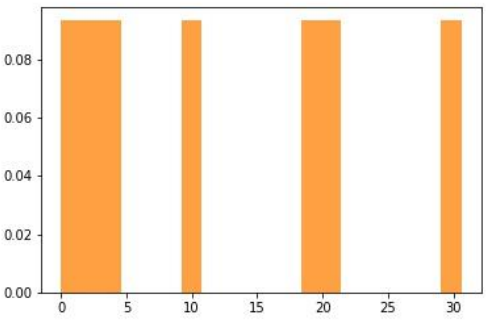

图 12 微博用户“_老番茄_”转发微博情感值频数分布直方图

对上述表格和频数分布直方图进行分析，微博用户“TFBOYS-易烊千玺”的微博情感极大部分(87.32%)都是积极向，小部分(8.70%)消极向的微博情感得分值都基本上分布在(-10, 0)区间内，情感均值为 8.3861，属于正面，情感方差为 81.3360。积极向的微博情感均值为 9.8108，方差为 75.931，消极向的微博情感均值为-2.0074，情感方差为 10.7798，可见消极向的微博聚集程度比积极向微博高。用户原创微博中，情感均值为 8.7066，情感方差为 90.5244，用户转发微博中，情感均值为 7.8559，方差为 65.6889。

对上述表格和频数分布直方图进行分析，微博用户“_老番茄_”的微博情感与正态分布较为拟合，大部分情感得分分布在(-10,10)区间，64 条原创微博的情感分布同样与正态分布较为拟合，7 条转发微博情感都是积极向。

表 3 为在“TFBOYS-易烊千玺”和“_老番茄_”两个用户微博数据集上的情感分析准确率分析：

表 5 基于情感词典的微博用户情感分析结果

数据集	微博总条数	原创条数	转发条数	积极条数	消极条数	正确条数	总准确率	原创准确率	转发准确率	积极准确率	消极准确率
TFBOYS-易烊千玺	828	516	312	811	17	740	0.8937	0.8876	0.9038	0.8915	1.000
老番茄	71	64	7	39	32	55	0.7746	0.7656	0.8571	0.8205	0.7188

可以看出，基于情感词典的用户情感分析在两个数据集上均有较高的正确率，但是，由于情感词典不能很好地分析出语义关系，微博文本的情感表现形式随着时代发展日新月异，情感词典也不能囊括所有情感词、出现的新词无法及时补充进情感词典，微博文本中还含有大量的广告消息，进行情感分析时也忽略了句子的标点符号等，使用情感词典对微博用户进行分析还是有很大的局限性。

5 总结

本次课程项目设计从选题、分析、设计到最后的实施、整理、完成设计报告，前前后后花了 8 天时间，从数据爬取、预处理到数据分析，每一步都有很大的挑战性。由于是第一次接触自然语言处理，本想用 word2vec 将文本向量化，借助机器学习相关知识来完成此次设计，无奈悟性不够，遂转用比较容易实现的基于情感词典的情感分析方法，花了大量的时间在数据爬取和预处理中。课程设计也算是把一学期以来在 Python 课上学到的知识加以实际应用，爬取数据中用 requests 库进行抓取、xpath 和 re 进行解析，数据分析时导入 wordcloud 模块绘制词云图，导入 numpy 和 matplotlib 进行数据分析和可视化展示，也用到了许多文件操作。虽然做的时候很累，遇到了很多问题，但是按照项目需求和数据科学的流程一步步走下来，也是一次很有意义的成长。

参考文献

[1] TFBOYS-易烊千玺的微博正文[2020-06-10] https://m.weibo.cn/u/3623353053

[2] _老番茄_的微博正文[2020-06-12] https://m.weibo.cn/u/3708072513

[3] BosonNLP 情感词典[2020-06-10] https://bosonnlp.com/dev/resource

[4] 《知网》情感分析用词语集（beta 版）[2020-06-10] http://www.keenage.com/download/sentiment.rar

[5] 停用词词典 [2020-06-10] https://github.com/isnowfy/snownlp/blob/master/snownlp/normal/stopwords.txt

[6] 否定词词典 [2020-6-10] https://kexue.fm/usr/uploads/2017/09/1922797046.zip

[7] 汉语词典 emoji 对照表 [2020-06-10] https://hanyucidian.18dao.cn/emoji