词性分析制作说明文档

1.1　项目概述与任务分工

黄樾	模型搭建与训练
郑佳楠	中文分词、文档撰写
孙浩安	中文分词、情感值可视化
吴书培	情感词典优化、文档撰写

随着 WEB2.0 时代的到来，通过在线社交媒体来搜集用户观点变得越来越普遍，推特、脸书、微博等平台已成为与读者分享想法和观点的有效渠道.这些信息获取的便利性大大吸引了那些观点挖掘者，并被用于商务、社会、教育和娱乐等特定目的。在此背景下，情感分析作为自然语言处理中的一个研究领域越来越受到学术界和工业界的关注，该领域中的情感分类、情感检测、情感预测等也成为近年来研究的热点。本项目按照图 1 所示流程图推进，情感分析过程中的数据流图如图 2 所示。

图 1 流程图

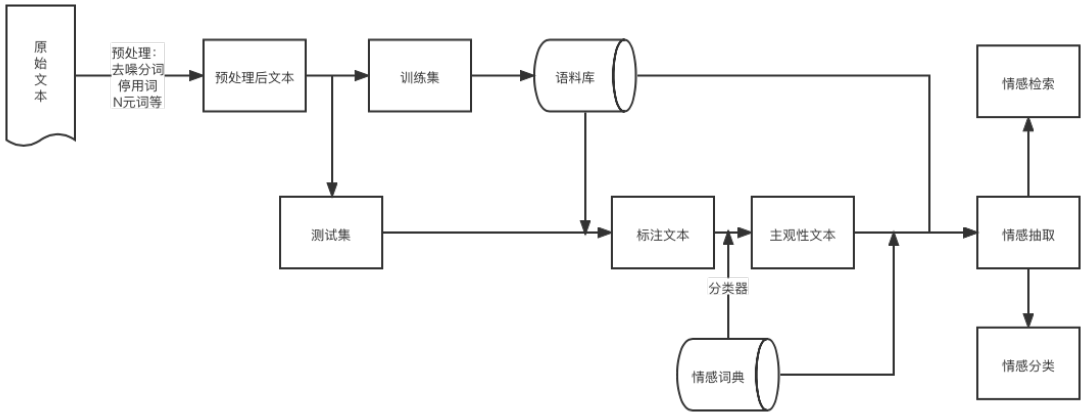

图 2 情感分析数据流图

1.2 情感词典

情感词典的构建是情感分类的前提和基础，目前在实际使用中，可将其归为 4 类:通用情感词、程度副词、否定词、领域词。目前国内外，情感词典的构建方法主要是利用已有电子词典扩展生成情感词典。英文方面主要是基于对英文词典 Word Net 的扩充,在手工建立种子形容词词汇表的基础上，利用 World Net 中词间的同义和近义关系可以判断情感词的情感倾向，并以此来判断观点的情感极性。中文方面则主要是对知网 Hownet 的扩充，可以利用语义相似度计算方法计算词语与基准情感词集的语义相似度，以此推断该词语的情感倾向。此外，还可以建立专门的领域词典，例如本项目中针对餐饮业评论这一场景补充用户词典，以提高情感分类的准确性。

购物领域情感词典构建：评论语料情感关键词+基础情感词典+手工添加情感词。训练集中的评论语料情感关键词全部提取出来作为最基本的情感词典情感词。此时我们可以通过人工挑选一些基础情感词典中符合购物领域情感词和人工添加情感词两种途径扩展我们的最基本情感词典。

基础情感词典：基础情感词典包括了一些被广泛认同的情感词，比如“好”，“漂亮”，“差”，“烂”这些词。将知网（Hownet）情感词典与台湾大学简体中文情感极性词典整合得到基础情感词典。但是这些基础情感词典有很多词，需要我们人工筛选判断并手动删除。

改进点：

1.拓展情感词典，拓展情感词典其实就是把基础情感词典通过同义词词典找到情感词的同义词，这样就拓展了基础情感词典。

2.人工添加情感词：从训练集中人工挑选出一些未记录在情感关键词中的情感词，并找一些购物领域评论相关的数据集，人工标注一些情感关键词并加入到我们的情感词典。基于情感词典匹配的情感分析，我们首先需要对测试数据集进行分词，然后与情感词典进行匹配，此时为了加快效率，我们通常将分词后的数据进行去停用词，停用词典的构建可以通过一些常用的停用词典，通过合并并人工筛选的方式构建。这样将一些无关简要的词停用掉。

1.3 分词

分词是将句子、段落、文章这种长文本，分解为以字词为单位的数据结构，方便后续的处理分析工作。英文有天然的空格作为分隔符，但是中文没有，所以如何切分是一个难点。再加上中文里一词多意的情况非常多，导致很容易出现歧义，信息爆炸的时代新词产生特别频繁，对新词的识别也是一大难处。

基于结巴分词的情感词典建立

一：结巴分词简介

1. 支持三种分词模式：

• 精确模式，试图将句子最精确地切开，适合文本分析；

• 全模式，把句子中所有的可以成词的词语都扫描出来，速度非常快，但是不能解决歧义；

• 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

2. 支持繁体分词

3. 支持自定义词典

主要功能：

1. 分词
2. 添加自定义词典
3. 关键字提取
4. 词性标注

我们同时还利用了百度开源 LAC 进行分词操作：main.py 中函数 cut(line)读入一行句子，将其中非中文字符去除后利用 LAC 分词，同样用生成由分好的词构成的 list 供后续处理。

1.4 主题词匹配与情感值获取

处理测试集时，对评论进行切词时引入前面建好的用户词典，然后依次判断切分的词是否在训练集中的正面词、中性词、负面词列表中，若在则标注其为情感词，并加上相应的标签。此时情感词已经获取到，在通过情感词推出其主题，过程是：情感词之前带有名词特点的词一般为主题词，根据这个原则寻找主题词，基于上述仍未找到主题词，则主题词为NULL。该过程流程图如图3所示。

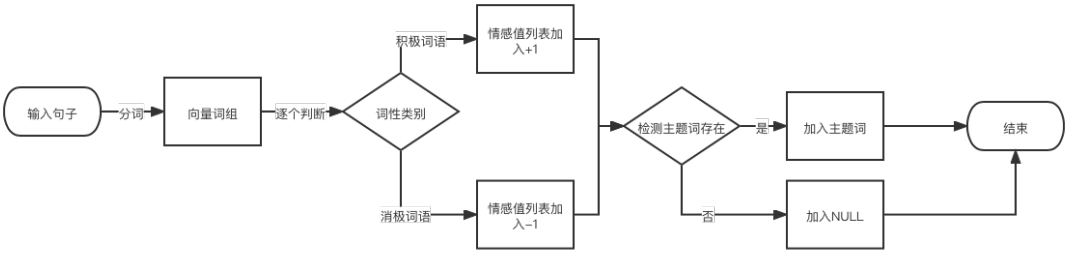

图 3 情感值获取流程图

处理结果如图 4 所示，D 列为情感词，C 列为对应主题词，而 E 列为对应情感值。

	A	B	C	D	E	F	G	H
1	评论内容	来源						
2	早就想吃这A	平台	土豆丝;NULL;黄喉;牛肉;	太大;入味;好吃;很嫩;	-1;1;1;1;			
3	真的觉得sir A	平台	早餐;地方;NULL;NULL;工具;NULL;蔷条;黑胡椒;	最好;不多;可以;很棒;专业;完美;好吃;棒;	1;0;1;1;1;1;1;1;			
4	N坊的西班 A	平台	店面;NULL;NULL;NULL;NULL;	小小的;好;喜歌;可以;五星;	-1;1;1;1;			
5	?? Shake S	A平台	世界;汉堡;NULL;味道;直饮水;	好吃;贵;很大;还不错;可以;	1;-1;1;1;1;			
6	大名鼎鼎的	A平台	NULL;氛围;风情;汉堡;满口;香;	大名鼎鼎;很好;美味;自然;香;喜欢;	1;1;1;1;1;1;			
7	美阿姨服务	A平台	服务态度;NULL;	非常好;和善;	1;1;			
8	菜品分量是	A平台	分量;辣子鸡;NULL;NULL;NULL;阿姨;服务;NULL;	好大;很香;入味;惊艳;很多;和蔼可亲;到位;赞;	-1;1;1;1;1;1;1;1;			
9	??环境：据	A平台	环境;店里;NULL;NULL;NULL;酱;NULL;	干净;很多;太热;惬意;满足;好吃;普通;	1;1;-1;1;1;0;			
10	味道非常好	A平台	味道;价格;NULL	非常好;偏贵;推荐;	1;-1;1;			
11	[服务铃]服	A平台	份量;服务;NULL;热情;	很好;周到;热情;满分;	1;1;1;1;			
12	终于吃到了	A平台	月饼;味道;外皮;	值得;不错;酥软;	1;1;1;			
13	之前一直在	A平台	NULL;猪脚;道菜;NULL;NULL;NULL;	棒;好吃;最爱;喜欢;卡;遣憾;	1;1;1;1;-1;-1;-1;			
14	田中猪子设	A平台	猪子;代表;地理位置;	很棒;不好吃;不错;	1;-1;1;			
15	菜品量非常	A平台	NULL;米遥;服务;体验;用餐;	推荐;好吸;非常好;用心;满意;	1;1;1;1;1;			
16	久闻大名的	A平台	大城市;饭点;下单;快餐店;NULL;NULL;味道;NULL;汉堡;耐受力;	幸运;容易;一般;迅速;贵;不便宜;惊艳;喜欢;好吃;极高;	-1;1;0;1;-1;-1;1;1;1;-1;			
17	[薄荷]环境	A平台	NULL;NULL;上菜;味道;酱汁;个头;馅肉;NULL;NULL;NULL;凉菜;感觉;	新鲜;不理;快;还不错;饱满;挺大;很紧;最好;推荐;不贵;普	1;-1;1;1;1;1;-1;1;1;1;1;0;0;1;1;1;1;1;-1;0;			
18	这一月一月	A平台	NULL;歌夏;NULL;环境;花园;下午茶;NULL;	快;好;低调;亮点;凉快;惬意;漂亮;	1;1;0;1;1;1;1;			
19	1「呐油鳝	A平台	NULL;NULL;NULL;巨;	推荐;厚;硬;难吃;	1;0;-1;-1;			
20	量大、味道	A平台	味道;冰粉;阿姨;	好;好吃;热情;	1;1;1;			
21	每次点评都	A平台	NULL;鸡汤;牛肉;火锅;价格;NULL;	好;完美;很嫩;最爱;精贵;值得;	1;1;1;1;-1;1;			
22	[薄荷]环境	A平台	NULL;人位;上菜;NULL;NULL;NULL;热干面;个头;味道;感觉;蒜泥;NU	正好;可以;很快;经典;很香;干;还不错;很大;奇怪;油腻;很	1;1;1;1;1;-1;1;1;1;-1;-1;1;1;1;1;0;-1;			
23	在新天地附	A平台	NULL;品种;NULL;NULL;汉堡;NULL;NULL;口感;	简约;并不多;好;好吃;不错;棒;赞;丰富;	1;-1;1;1;1;1;1;1;			
24	和武汉吃至	A平台	辣度;虾;NULL;NULL;毛豆;武汉;	符合;干净;新鲜;最好;最爱;不喜欢;	1;1;1;1;1;-1;			
25	朋友请客选	A平台	NULL;服务;NULL;菜码;菜;NULL;	惊喜;很好;赞;好大;好吃;还可以;	1;1;1;-1;1;1;			
26	之前一直看	A平台	汉堡;NULL;价格;觉得;NULL;味道;觉得;	好吃;花了;蛮高;很失望;难吃;还可以;失望;	1;-1;1;-1;-1;1;-1;			
27	好吃，量大	A平台	NULL;服务;	好吃;热情;	1;1;			
28	非常喜欢的	A平台	NULL;	非常喜欢;	1;			
29	打卡网红餐	A平台	NULL;辣味;菜品;菜品;味道;	花了;少;好吃;量爱;很好;	-1;-1;1;1;1;			
30	老老上海了	A平台	味道;本帮菜;水;	经典;好吃;推荐;	1;1;1;			
31	八佰伴旁边	A平台	NULL;环境;位置;量;鸡肉;NULL;NULL;性价比;	好吃;还可以;好;很大;很多;脆;甜甜的;高;	1;1;1;1;1;1;1;1;			
32	之前一直吃	A平台	定位;NULL;NULL;	可以;好;免费;	1;1;1;			
33	霸王餐抽到	A平台	料;NULL;	足;美;	1;1;			
34	到达上海，	A平台	排队;NULL;安排;NULL;NULL;NULL;NULL;NULL;NULL;外地;NULL;	随便;快;满意;推荐;不错;理解;可以;吃;不消;不适合;值重	-1;1;1;1;1;1;1;-1;-1;-1;1;			
35	口味一般，	A平台	口味;长沙;	一般;入味;	0;1;			
36	这家shake	A平台	环境;座位;感觉;服务;NULL;	一般般;不够多;一般;不错;快;	0;-1;0;1;1;			
37	周末晚上去	A平台	流动;NULL;NULL;NULL;	蛮大;清爽;很多;厚;	1;1;1;0;			
38	非常红火的	A平台	等位;速度;NULL;NULL;饭;NULL;	很多;不算快;中规中矩;辣辣的;好;好吃;	1;-1;0;-1;1;1;			
39	公司附近的	A平台	份量;女生;地道;	很大;贴心;正宗;	1;1;1;			

result1

图 4 情感分析结果

1.5 可视化

对于主题词情感值可视化：theme_words_emotion_visualization.py

使用 matplotlib.pyplot 包输出主题词—情感值的柱状图，支持显示集中每行句子中的主题词积累的总情感值和平均情感值，支持显示情感值结果为正数(mode 参数为"positive")或负数(mode 参数为“negative”)的排名靠前的主题词。

主要内容有：

全局变量：设置正面/负面倾向主题词出现次数，只有高于此次数的数据才会被统计，设置显示情感值前多少个的正面/负面倾向主题词。

函数：

getSumThemeEmotionList(path, mode="positive")：利用读取 main.py 已经生成的 result 文件统计主题词的总情感值，并进行排序，返回以两个列表为元素的列表，其中第一个列表储存主题词，第一个列表储存主题词对应总情感值。

getAvgThemeEmotionList(path, mode="positive")：统计主题词平均情感值返回以两个列表为元素的列表，其中第一个列表储存主题词，第二个列表储存主题词对应平均情感值。

visualization(theme_emotion_list) ：利用以上两个函数生成的列表输出柱状图表，支持窗口大小、字体大小、角度变化，含数值标注。

如图 5 为可视化结果，可视化的柱状图表横坐标为主题词，纵坐标为对应情感值，颜色同情感值数值大小相关，越接近绿色表示该主题词通常与积极的情绪相关，红色则代表消极。

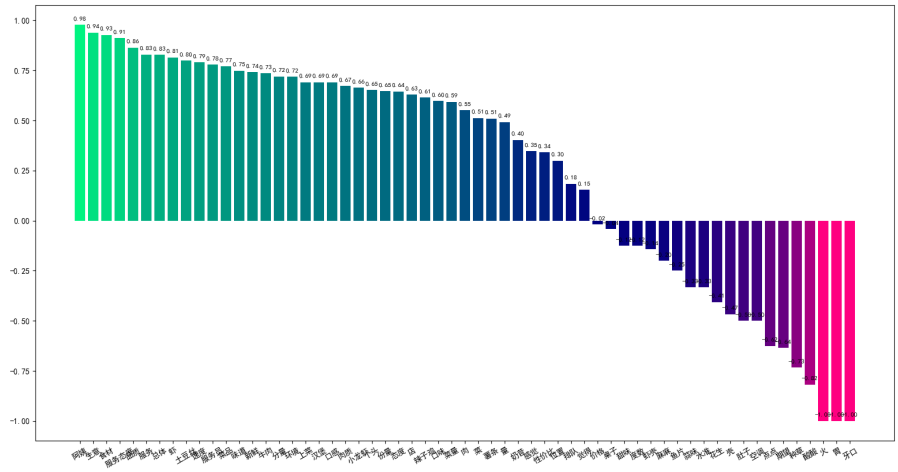

图 5 可视化结果

1.6 不足与改进

利用情感词典来判断一个句子的情感是有着明显不足的。中文有着丰富的语义表达，很多情感都是隐含的，比如：”我昨天吃了这道菜，今天就拉肚子了“。这句话没有一个情感词，但表达的是消极的情绪。还有各种事正话反说的句子，比如：“你说这里的菜很好吃，我只能呵呵了“。如果用词典匹配，有”好吃“这个积极词，但这句话表达的绝非积极的情绪。这里就需要更高级复杂的处理方式，要更深入句子的句法、语法了，我们只实现了最简单最基础的一些规则。此外，词语的情感值在随着社会事件与背景环境发生改变，情感词典无法抓取到这些变化。