基于微博的情绪分类

一、研究内容

近数十年来，人工智能领域发展迅猛，总的来说，包含三个阶段：计算智能、感知智能和认知智能。在计算智能阶段，人类早已不能匹敌计算机，该阶段的计算机配置有超大的存储容量和超高的计算速度，能存会算。在感知智能阶段，国内外学者们致力于研究如何让机器模仿人类的五种感觉器官，即视觉、听觉、嗅觉、味觉和触觉，得益于深度学习理论日趋完善，在该阶段学术界和工业界的成果斐然，尤其在以语音识别、图像识别为代表的感知任务上已经可以和普通人相媲美，甚至在某些方面赶超普通人。在认知智能阶段，机器处理的对象是人类语言（词、短语、句子、篇章）、思维等高层抽象的东西，需要机器能够思考、理解，具有情感以及自我学习的能力，也就是真正具备人的智能。很显然，与感知智能阶段的任务相比，认知智能阶段的任务更具挑战性。以苹果公司的 Siri 为代表的聊天机器人已经渗透到普通人的日常生活中，孤独的时候拿起手机可以和 Siri 聊聊天、解解闷，虽然Siri 会卖个萌，也懂一些小幽默，但距离真正的共情、理解人类的情感还是有非常遥远的距离。情感计算旨在通过赋予机器识别、理解、表达和适应人类情感的能力来构建和谐的人机交互环境，并助力机器拥有更高、更全面的智能。因此，情感计算在认知智能阶段非常重要，是当代认知科学的一大研究热点和难点。

语言是将人类区别于其他物种的最显著特征，是人类最重要的交际工具，而文字是语言的视觉形式，突破了各个人种、口音的限制，将人类文明传承下来的最重要的载体。自然语言处理（Natural Language Processing, NLP）就是一个旨在研究如何使用计算机技术来处理、理解以及运用人类语言的各种理论和方法的领域以自然语言处理为代表的认知技术是人工智能的核心技术之一，是促进人工智能不断进步的阶梯。

文本情感计算作为自然语言处理领域的一个重要分支，其基本目标是从给定文本中识别出用户的观点信息，并分析该观点中所蕴含的态度、评价或情绪。文本情感计算的应用前景非常广泛，无论个人、商家还是政府单位，都在逐步利用这一项研究的成果来辅助决策。对于个人而言，假如你有意在天猫商城上购买一款笔记本电脑，那么你可以先搜集其他用户对该笔记本电脑的评论文本，然后利用文本情感计算方法分析得到其他用户对该笔记本电脑的总体情感倾向以及各个方面的情感倾向，最后你就可以根据自己的需求来决定是否购买这款笔记本电脑了。站在商家的角度来说，根据一个用户在各大社会媒体或电子商务平台上发布的帖子和评论，我们可以使用机器学习算法从中挖掘出该用户关注的主题信息和观点喜好，生成该用户的情感画像，然后依据这些对用户的兴趣刻画，商家能够及时知晓用户对其产品和服务的评价和喜好，并有效地向用户进行个性化产品和服务推荐。对于政府而言，当今各国政府如果想要推行某项政策，通常会先选择在某些地方政府进行试点，然后搜集公众对于该项政策的观点和看法，接着做政策舆情分析，得到公众对于该项政策的总体情感倾向，最后迅速做出调整，以期最大化地服务于人民，引领整个社会、经济和政治环境朝着良好的方向发展。此外，根据上映头几天的用户观影评论来预测新上映电影的票房，根据民众的 twitter 言论来预测政治选举结果，根据股票行情评论来预测股票走势等等，文本情感计算研究在普通人的日常生活中的应用比比皆是。当然，文本情感计算还可以应用于其他自然语言处理任务，使得融入情感因素的智能系统更加人性化。例如，融入情感因素的智能谈判系统往

本实验中，通过对用户的情感进行分析，将情绪分为正向和负向两类。其中label值为1表示正向评论，label值为0表示负向评论

Table 1实验数据示例

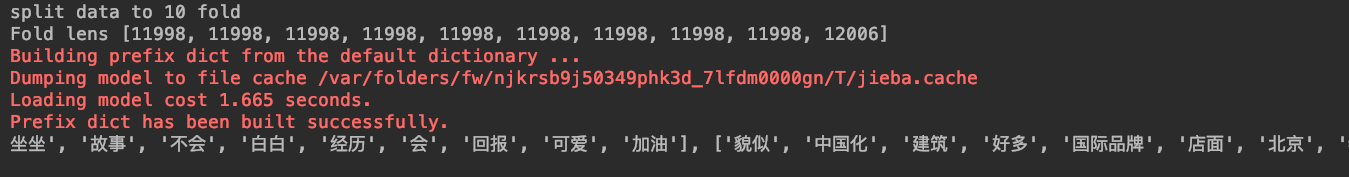
|  |  |
| --- | --- |
| label | 微博内容 |
| 0 | 太过分了@Rexzhenghao //@Janie\_Zhang:招行最近负面新闻越来越多呀... |
| 0 | 希望你?得好?我本＂?肥血?史＂[晕][哈哈]@Pete三姑父 |
| 0 | 有点想参加????[偷?]想安排下时间再决定[抓狂]//@黑晶晶crystal: @细腿大羽... |
| 1 | [给力]感谢所有支持雯婕的芝麻！[爱你] |
| 1 | 2013最后一天，在新加坡开心度过，向所有的朋友们问声：新年快乐！2014年，我们会更好[调... |
| 0 | 大中午出门办事找错路，曝晒中。要多杯具有多杯具。[泪][泪][汗] |
| 0 | 马航还会否认吗？到底在隐瞒啥呢？[抓狂]//@头条新闻: 转发微博 |
| 0 | 克罗地亚球迷很爱放烟火！球又没进，就硝烟四起。[晕] |
| 1 | [抱抱]福芦 TangRoulou 吉祥书 8.8折优惠 >>> http://t.cn/z... |
| 1 | 回复@钱旭明QXM:[嘻嘻][嘻嘻] //@钱旭明QXM:杨大哥[good][good][g... |
| 1 | 人家这脸长的!!!!!![哈哈] |
| 1 | 这个价不算高，和一天内训相比相差无几。。[哈哈]//@博通传媒v: 6个月！一个月工资1万，... |
| 0 | 终于收工啦，脚丫子快冻掉了[泪][泪][泪] |
| 0 | 我决定从今天开始我想吃什么就去吃什么，一个人吃也无所谓，重点是不要因为别人的意见委屈了自己[... |
| 1 | 飘雪的北京 需要双份早餐.......//@美食天下: [哈哈]//@王淼Margay: 屁... |
| 1 | [耶]，这个太赞了，生活大爆炸第六季马上要出啦[鼓掌] //@-郑瑜-:这个不错 //@经典... |

二、实验方法

1. 文本分词

词是最小的能够独立活动的有意义的语言成分，英文单词之间是以空格作为自然分界符的，而汉语是以字为基本的书写单位，词语之间没有明显的区分标记，因此，在使用该微博文本数据之前，需要对其进行分词，本实验使用jieba分词工具对文本进行分词。

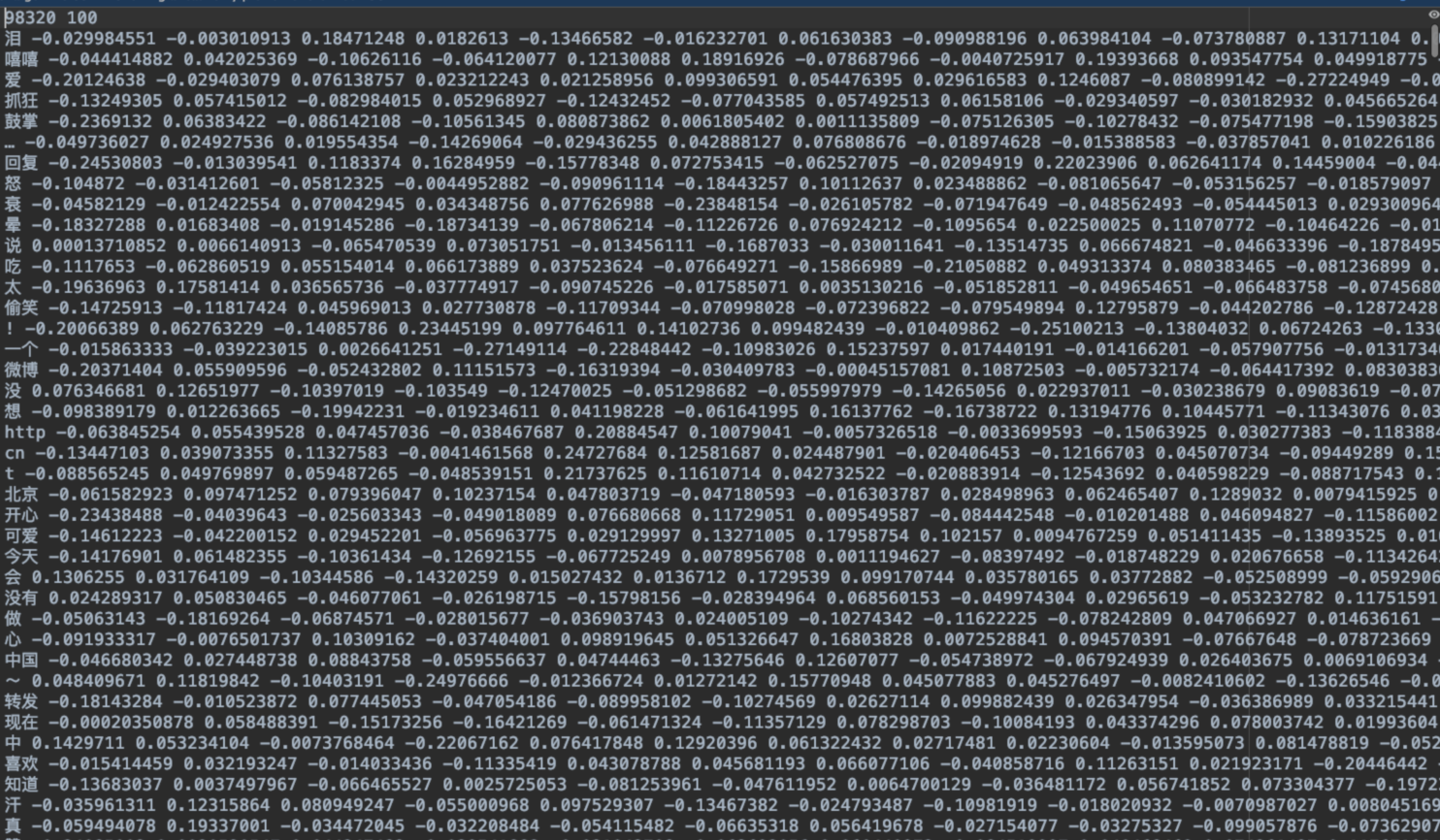
Figure 1 分词结果示例



1. word2vec/词向量

word2vec，即词向量，将一个词用一个向量来表示。是2013年Google提出的。word2vec工具主要包含两个模型：跳字模型（skip-gram）和连续词袋模型（continuous bag of words，简称CBOW），以及两种高效训练的方法：负采样（negative sampling）和层序softmax（hierarchical softmax）。word2vec词向量可以较好地表达不同词之间的相似和类比关系。word2vec是一个NLP工具，它可以将所有的词向量化，这样词与词之间就可以定量的去度量他们之间的关系，挖掘词之间的联系。

Figure 2 词向量表示

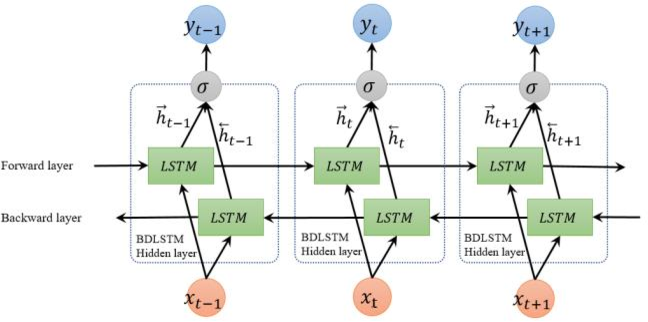


1. Bi-LSTM模型

实验使用Bi-LSTM模型对文本进行分类，LSTM的全称是Long Short-Term Memory，它是RNN（Recurrent Neural Network）的一种。LSTM由于其设计的特点，非常适合用于对时序数据的建模，如文本数据。Bi-LSTM是Bi-directional Long Short-Term Memory的缩写，是由前向LSTM与后向LSTM组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。将词的表示组合成句子的表示，可以采用相加的方法，即将所有词的表示进行加和，或者取平均等方法，但是这些方法没有考虑到词语在句子中前后顺序。使用LSTM模型可以更好的捕捉到较长距离的依赖关系。因为LSTM通过训练过程可以学到记忆哪些信息和遗忘哪些信息。

但是利用LSTM对句子进行建模还存在一个问题：无法编码从后到前的信息。在更细粒度的分类时，如对于强程度的褒义、弱程度的褒义、中性、弱程度的贬义、强程度的贬义的五分类任务需要注意情感词、程度词、否定词之间的交互。举一个例子，“这个餐厅脏得不行，没有隔壁好”，这里的“不行”是对“脏”的程度的一种修饰，通过Bi-LSTM可以更好的捕捉双向的语义依赖，拓宽模型的视野域。

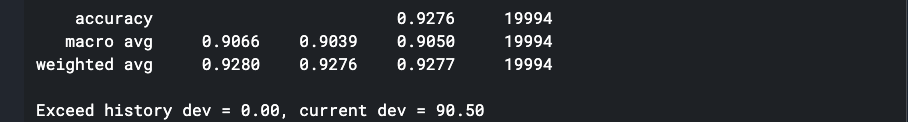
Figure 3 Bi-LSTM



1. 模型详解

根据词向量和数据语料建立词表。在处理句子时，对于每一个句子以一个最长的句子的长度作为固定长度，其它的句子长度不足的进行填充。使用两层双向LSTM，输入一个句子后产生的输出，取中间结果作为输出，输入attention层，进行加权平均得到一个句子表示。将句子的向量表示，输入到以交叉熵作为损失函数的线性分类器中，进行分类，得到最终的结果。

1. 实验结果



1. 课程感悟

学习了《机器翻译》这门课，课堂上的学习和实践，使我对自然语言处理有了一个宏观上的了解，课程的内容可谓丰富多彩，开阔了我的眼界和思维。我本身的研究方向是时间序列数据处理与自然语言同为序列化数据，课堂上学习到的模型，也对我以后的研究有很大的帮助。