湖南大學

HUNAN UNIVERSITY

微博评论情感分析

学号：S　姓名：zxy

目录

1. 引言.....1
2. 问题定义.....1
3. 研究思路及算法.....1
    3.1 常规方法预处理步骤.....2
        3.1.1 构造词特征.....2
        3.1.2 特征降维——jieba 分词.....2
        3.1.3 特征表示——jieba 分词.....3
    3.2 Bi-LSTM 算法.....3
4. 实验评估与分析.....4
    4.1 数据集.....4
    4.2 模型基线.....5
        4.2.1 BernoulliNB.....5
        4.2.2 MultinomiaNB.....5
        4.2.3 LogisticRegression.....5
        4.2.4 SVC 与 NuSVM.....5
        4.2.5 LinearSVC.....5
    4.3 评价标准.....5
        4.3.1 准确率(Accuracy).....6
        4.3.2 召回率(Recall).....6
        4.3.3 精准率(Precision).....6
        4.3.4 特异性(Specificity).....6
        4.3.5 F1-Score.....7
        4.3.6 AUC 曲线.....7
    4.4 结果与分析.....8
5. 相关工作.....10
6. 未来展望.....10
7. 总结.....11
参考文献.....11

1. 引言

微博是大众交换信息的一个重要平台，微博评论的情感分析可以获取大众对某件事物的看法，其中潜藏着商业价值，并且情感分析也可促进社会稳定（如网络谣言等）。情感分析又称意见挖掘，主要是对带有感情色彩的主观性文本进行分析、处理、归纳然后进行推理的过程。例如在网购过程中关于商品的评论，通过分析历史评论的情感极性，可以让消费者快速了解商品的质量，从而优化购买的决策，同时卖家也可以从评论中得知商品的问题所在，更精确地提升的服务和改善商品的质量。

学者在文本挖掘挖掘展开了很多的研究，微博，作为一个国内较为火热的社交平台，其涉及的内容十分广发，如娱乐、影视、体育等，不同的内容有着不同的研究，本文结合课上和课后所学知识，运用现在的微博情感分析方法，开展本次实验。实验表明LinearSVM和Bi-LSTM的效果比较好，准确率达到96%以上。

虽然LinearSVM比较简单[1]，但是调优LinearSVM的训练这个过程是相当有启发性的事情，LinearSVM每次只取使得损失函数极大的一个样本进行梯度下降，这就导致模型在某个地方可能来来回回都只受那么几个样本的影响，加上对测试集的未知性，最终还是采用Bi-LSTM进行情感分类。

2. 问题定义

本次项目的实验数据集是微博评论，每条记录有两个维度，分别为评论内容和已经标注好的情感标签。情感标签值为0或1，0代表负面情绪，1代表正面情绪。项目所使用的模型主要任务是学习数据集中的情感分类情况，当有新的数据输入到模型中时，模型能够判断出这些新数据的情感值并输出情感标签即1或者0。

3. 研究思路及算法

这里主要分为两大类来实现微博情感分类，首先尝试实现早期学者解决情感分类问题所运用的数学概率统计的方法，这些方法在实现前都要经过词特征的构建、特征降维和特征表示；后期由于深度学习的兴起，许多学者开始用神经网络的思想来进行情感分类[2]，一开始将 CNN 运用到情感分类中有多成果，而 CNN 无法解决长距离依赖的问题，在此基础上，Hochreiter & Schmidhuber (1997)[3]提出提出了长短期记忆网络(Long Short Term Memory Network，LSTM)，如下图 1 所示，LSTM 可以解决训练过程中的梯度消失和梯度爆炸问题[4]，简单来说就是在更长序列中有更好的表现，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 BiLSTM 可以更好的捕捉双向的语义依赖。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，也由此它可以兼顾上下文，比 LSTM 效果好很多。

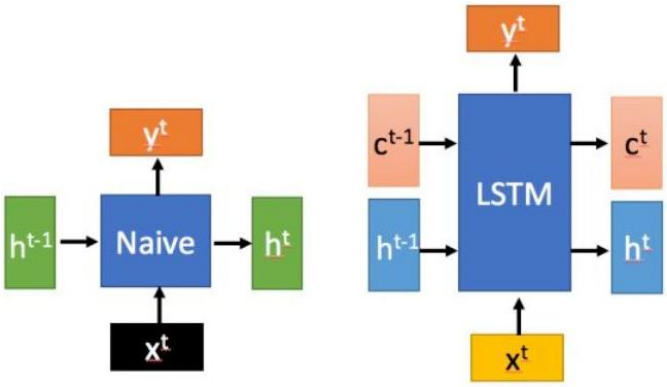

图 1 CNN 与 LSTM 的对比(左边为 CNN，输入是 x^t 和 h^{t-1}，输出是 y^t, h^t；右边是 LSTM，与 CNN 不同的是加入了 c^t，其是记忆细胞，负责储存和更新单词)

3.1 常规方法预处理步骤

3.1.1 构造词特征

评论是句子拼接在一起表达评论人情感的，词是组成句子的最小单位，处理文本时，首先要将文本分割成词，最常用的是jieba分词，分完词后，句子便由一个列表，里面包含着很多词。如下图2所示，而词语词之间是有关系的，例如，“吃”和“饭”这两个词搭配应是十分频繁的，因此，处理单个词作为评论文本特征以外，这里还考虑了双词搭配(Bigrams)，比如“手机 非常”，“非常 好用”，“好用!”这三个搭配作为分类的特征，以及单个词加双词的搭配。

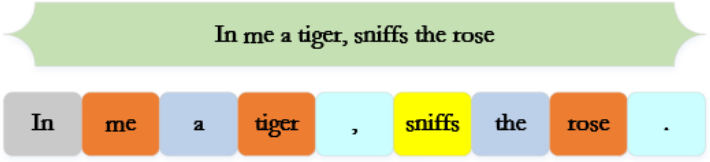

图2 分词例句。绿色表示原始句子，小格块表示分词后句子，不同颜色表示不同的词性

3.1.2 特征降维——jieba 分词

对于初始的文本，需要进行停用词过滤，然后对积极评论和消极评论分别进行词频的统计，接着对照情绪词袋库，对于包含在词库中的词，词的信息量等于

卡方统计量，对于不包含在词库中的词，词的信息量等于卡方统计量加上偏值，最后将所有词按照信息量倒序排序，自定义维度即选取信息量前n的词，如下图3所示：

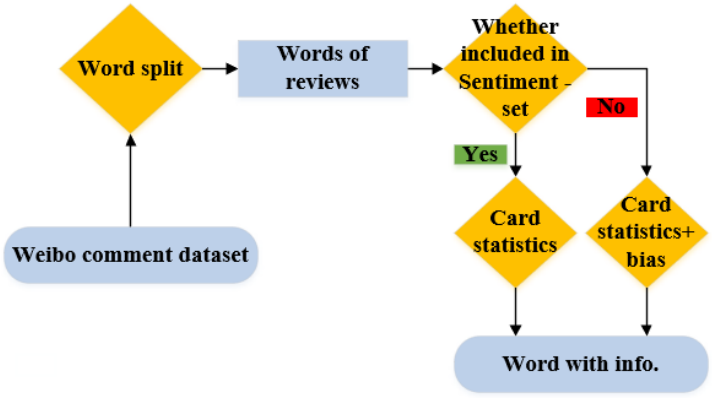

图3 卡方统计获取信息量前n的词。其中“Card”表示卡方统计，“sentiment-set”表示自定义情绪词集，“bias”为自定义的偏值

3.1.3 特征表示——jieba 分词

根据训练集中的标签，评论文本可以分成积极评论文本和消极评论文本，将这两种类型的文本分别在基于上述卡方统计后的词集进行标记即评论文本中是否包含信息量较大的词，最后依据标签，如下图4所示，把积极评论与消极评论标记上“pos”和“neg”。

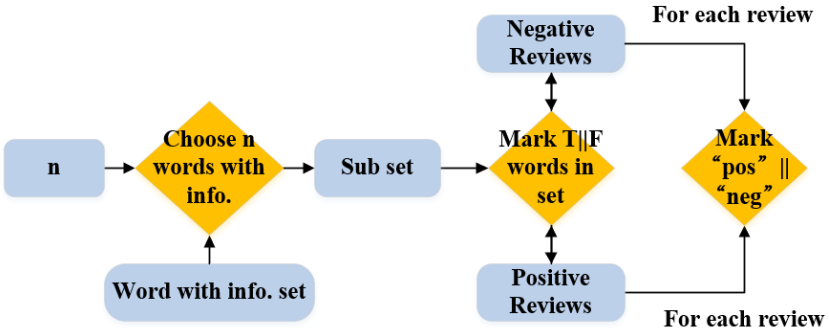

图4 根据卡方统计获取信息量前n的词和训练集的标签分步进行标记

3.2 Bi-LSTM 算法

LSTM通过门控状态来控制传输状态，记住需要长时间记忆的，忘记不重要的信息，但是RNN和LSTM都只能依据之前时刻的时序信息来预测下一时刻的输出，但在有些问题中，当前时刻的输出不仅和之前的状态有关，还可能和未来的状态有关系。比如预测一句话中缺失的单词不仅需要根据前文来判断，还需要考虑它后面的内容，真正做到基于上下文判断，而在情感分类中，需要根据词前词

后的上下文语境，才能更好地判断单词的情感极性，例如“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，这样更准确确定整条评论的情感状态。本次实验中算法流程图如下图5所示，模型最终输出一个概率值p，取值范围为[0,1]。

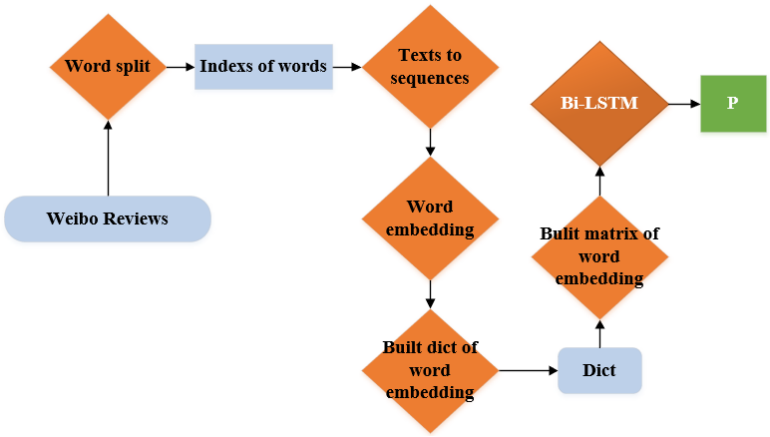

图5 运用Bi-LSTM进行情感分类

4. 实验评估与分析

4.1 数据集

本次实验的数据集是带有标签的微博评论数据，如下图 6 所示，评论数据条数为 107998，其中是 53995 条正面的评论即标签为 1，53993 条数据是负面的评论即标签为 0 的数据，数据分布较为均匀，因此采样时，不采用一些特别的处理手段。为了验证模型的效果，会将这个 train dataset 以 2/8 原则进行划分为 sub train set 和 val set。

表 1 微博评论数据集样例

Review	Label
你们这些大骗子，不是说不疼吗？[泪]	0
色总，您是大明星啊 //@辞小西:噗哈哈哈哈哈哈！！！[哈哈]哎呀我去啊，色总那羞涩的眼神...	1
反正比喵总漂亮的第二天都得。。。[嘻嘻] //@冯绍峰:能低调点嘛[挖鼻屎]	1
以后谁也不要请我喝可乐！我跟可乐犯冲。带狗出去没法进饭店就去 KFC 买了全家桶回家，还没到家三...	0
上次真应该囤点巧克力 听了你@chloe518 建议 中午火速买黑巧 希望下午暖和点吧 别...	0

4.2 模型基线

4.2.1 BernoulliNB

伯努利贝叶斯假设数据服从多元伯努利分布，并在此基础上应用朴素贝叶斯的训练和分类过程。多元伯努利分布简单来说，就是数据集中可以存在多个特征，但是每个特征都是二分类的，可以用布尔变量表示，即{0，1}或者{-1，1}等任意二分类组合，因此这个类要求将样本转换为二分类特征向量，如果数据本身不是二分类的，可以使用类中专门用来二值化的参数来改变数据。

4.2.2 MultinomiaNB

MultinomialNB主要用于离散特征分类，假定输入数据为计数数据（即每个特征代表某个对象的整数计数，比如一个单词在句子里出现的次数）。

伯努利贝朴素贝叶斯与多项式朴素贝叶斯非常相似，都用于处理文本分类数据，但是由于伯努利朴素贝叶斯是处理二项分布的，所以它更在意的是“存在是或否”，而不是“出现多少次”这样的次数或频率，这就是伯努利贝叶斯与多项式贝叶斯的根本性不同。在文本分类时，伯努利朴素贝叶斯可以使用单词出现向量（而不是单词计数向量）来训练分类器。文档较短的数据集上，伯努利朴素贝叶斯的效果会更好。

4.2.3 LogisticRegression

逻辑回归是一种广义线性回归模型，是Sigmoid函数归一化后的线性回归模型，常用来解决二元分类问题，可解释性强。它假设数据服从伯努利分布，通过梯度下降法对其损失函数（极大似然函数）求解，以达到数据二分类的目的。

4.2.4 SVC 与 NuSVM

支持向量分类数据拟合的时间复杂度是数据样本的二次方，这使得他很难扩展到10000个数据集，当输入是多类别时（SVM最初是处理二分类问题的），通过一对一的方案解决，当然也有别的解决办法。Nu支持向量分类。与SVC相似，但使用参数来控制SVM的数量。

4.2.5 LinearSVC

此次试验主要用到的是LinearSVC（Linear Support Vector Classification）分类器，线性支持向量分类，类似于SVC，但是其使用的核函数是”linear“上边介绍的两种是按照brf（径向基函数计算的，其实现也不是基于LIBSVM，所以它具有更大的灵活性在选择处罚和损失函数时，而且可以适应更大的数据集，他支持密集和稀疏的输入是通过一对一的方式解决。

4.3 评价标准

机器学习分类任务的常用评价指标：混淆矩阵（Confuse Matrix）、准确率

（Accuracy）、精确率（Precision）、召回率（Recall）、F1 Score、P-R曲线（Precision-Recall Curve）、ROC、AUC。针对一个二分类问题，即将实例分成正类（Positive）或负类（Negative），在实际分类中会出现以下四种情况。

（1）若一个实例是正类，并且被预测为正类，即为真正类TP(True Positive)

（2）若一个实例是正类，但是被预测为负类，即为假负类FN(False Negative)

（3）若一个实例是负类，但是被预测为正类，即为假正类FP(False Positive)

（4）若一个实例是负类，并且被预测为负类，即为真负类TN(True Negative)

混淆矩阵的每一行是样本的预测分类，，如下表2所示，每一列是样本的真实分类（反过来也可以）。

表2 混淆矩阵

True label ~~pre. label~~	Positive sample	Negative sample
Positive sample	TP	FN
Negative sample	FP	TN

4.3.1 准确率(Accuracy)

预测正确的样本数量占总量的百分比，具体的公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

4.3.2 召回率(Recall)

又称为查全率或者敏感性（Sensitivity），是针对原始样本而言的一个评价指标。在实际为正样本中，被预测为正样本所占的百分比。具体公式如下：

$$\text{Recall} = \frac{TP}{TP + FN}$$

尽量检测数据，不遗漏数据，所谓的宁肯错杀一千，不肯放过一个，分类阈值较低。

4.3.3 精准率(Precision)

又称为查准率，是针对预测结果而言的一个评价指标。在模型预测为正样本的结果中，真正是正样本所占的百分比，具体公式如下：

$$\text{Precision} = \frac{TP}{TP + FP}$$

精准率的含义就是在预测为正样本的结果中，有多少是准确的。这个指标比较谨慎，分类阈值较高。

4.3.4 特异性(Specificity)

真负类率（True Negative Rate, TNR），也称为特异性(Specificity)，刻画的是被分类器正确分类的负实例占所有负实例的比例。

$$TNR = \frac{TN}{FP + TN}$$

负正类率（false positive rate, FPR），也称为1-specificity，计算的是被分类器错认为正类的负实例占所有负实例的比例。

$$FPR = 1 - TNR = \frac{FP}{FP + TN}$$

4.3.5 F1-Score

单独使用准确率和召回率并不能很好的评估推荐系统，因为它们都依赖推荐列表，并且两者负相关，并且推荐结果越多，准确率相应减少越小，召回率增加。精确率和召回率是此消彼长的，如果要兼顾二者，就需要F1-Score

$$F1 = \frac{2 \times \text{Precision} \times \text{Re}\,call}{\text{Precision} + \text{Re}\,call}$$

4.3.6 AUC 曲线

以逻辑回归分类器为例[6]，其给出针对每个实例为正类的概率，那么通过设定一个阈值如0.6，概率大于等于0.6的为正类，小于0.6的为负类。对应的就可以算出一组(FPR,TPR)，在平面中得到对应坐标点。随着阈值的逐渐减小，越来越多的实例被划分为正类，但是这些正类中同样也掺杂着真正的负实例，即TPR和FPR会同时增大。阈值最大时，对应坐标点为(0,0),阈值最小时，对应坐标点(1,1)。如下面这幅图6中的(a)图中，实线为ROC曲线，线上每个点对应一个阈值。计算出ROC曲线下面的面积，就是AUC的值。

假设分类器的输出是样本属于正类的socre（置信度），则AUC的物理意义为任取一对（正、负）样本，正样本的score大于负样本的score的概率。AUC越大，分类器分类效果越好。

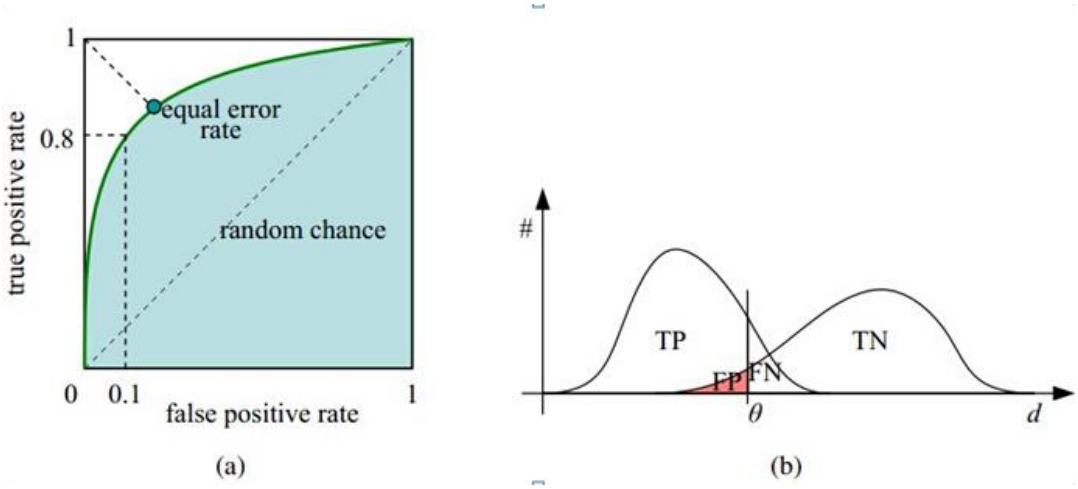

图6 ROC与AUC

4.4 结果与分析

特征提取的方法有很多，因此我们进行实验来探寻一个较为好的方法，如下图7所示，可以看出，当jieba_feature特征提取和词袋特征提取效果，在6个情感分类器中都比较稳定。

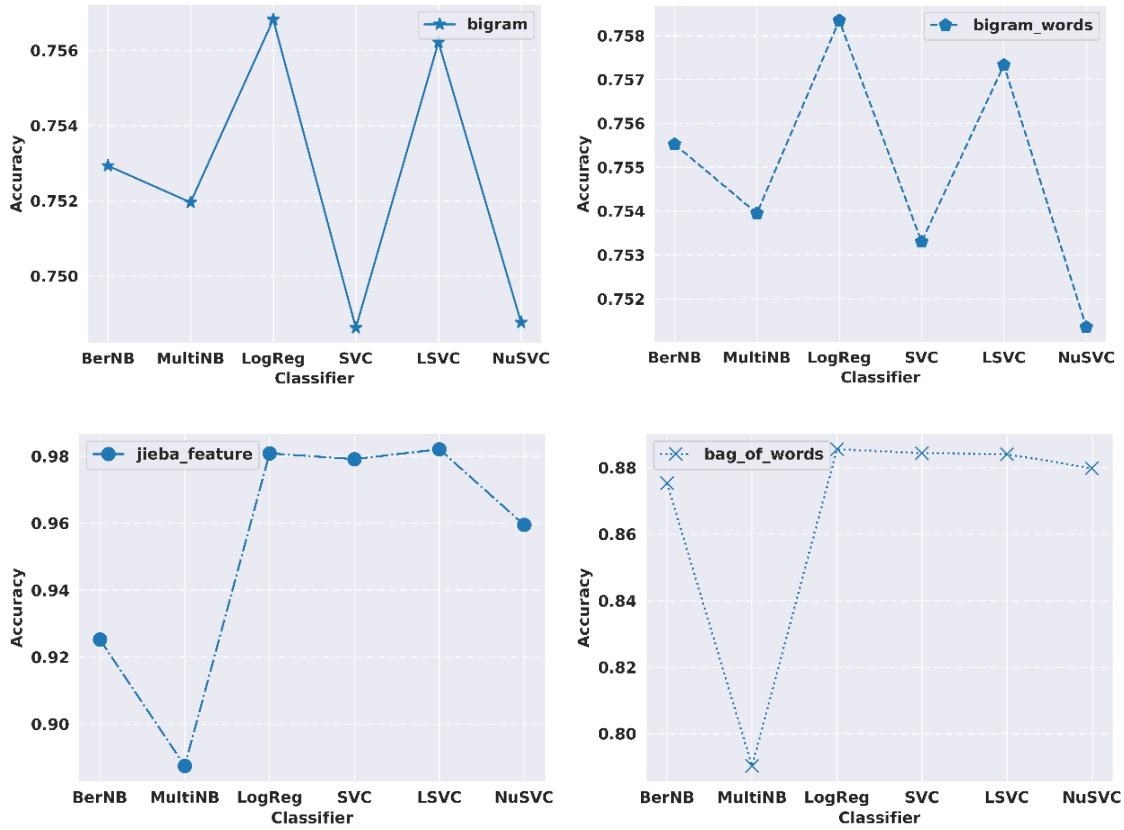

图7 词特征对分类器的影响。横坐标表示分别是BernoulliNB classifier、MultinomiaNB classifier、LogisticRegression classifier、SVC classifier、LinearSVC classifier、NuSVC classifier，这些分类器在bigram、double bigram、jieba feature extraction and bag of words下的准确率

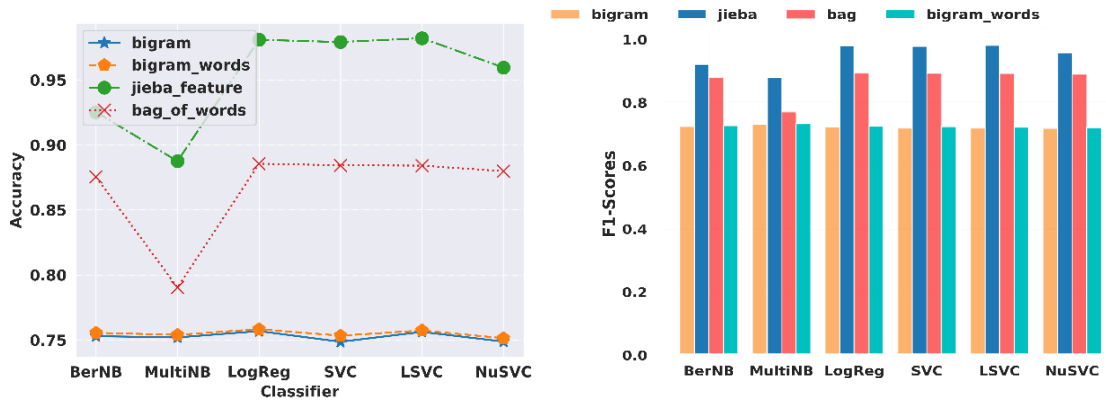

图8 特征提取与分类器之间的关系

进一步探究特征提取器与分类器的关系，如上图8所示，可以发现LinearSV-M、SVC和线性回归方法的效果比较好，准确率达到95%左右，F-Score值亦是这三个分类器表现较为出类拔萃。

下图9是Bi-LSTM模型在微博评论数据集上的分类效果，实验中总共用了6次迭代，可以看出模型的准确率已经很高了，模型损失也在第2、3次迭代后快速下降，模型在训练集和验证集上都比较稳定，一方面的原因在于本身数据标签比较精准，使得模型“学习效率比较高”。

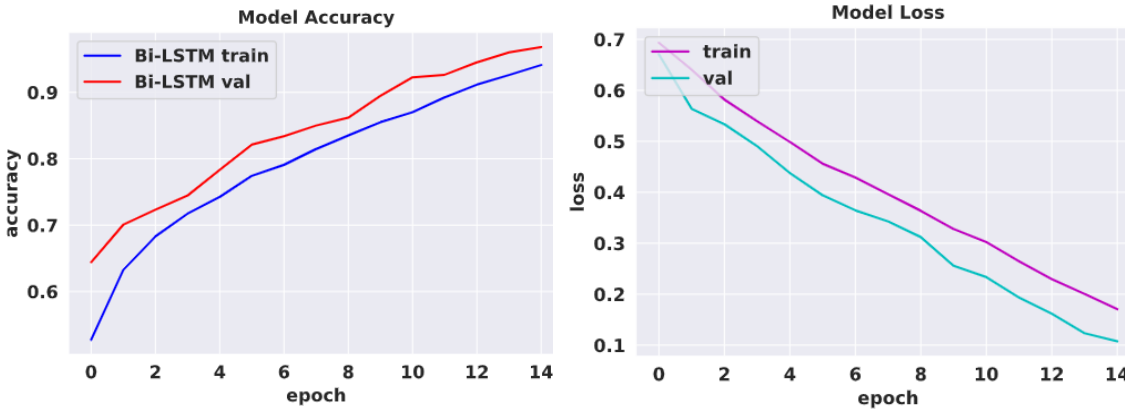

图9 Bi-LSTM情感分类器效果(左边为模型的准确率，右边为模型的损失)

通过AUC曲面面积和混淆矩阵，如图10-11所示，可以清晰地看出Bi-LSTM模型的优越性，其中F1-Score是96.78%，Recall是95.1%，精确率是98.52%，准确率是96.83%（可以多跑几次模型，效果可能更好，但是这种不尝试了）。

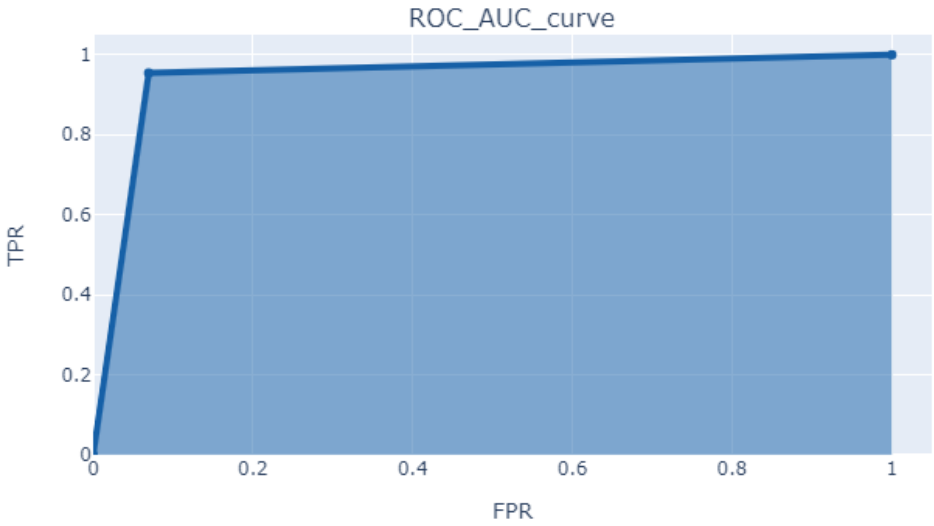

图10 词特征对分类器的影响。

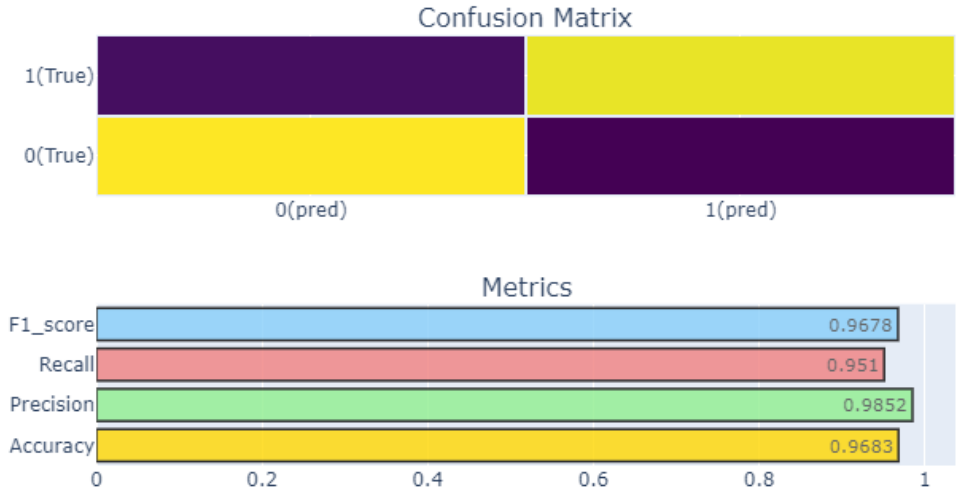

图11 词特征对分类器的影响。

5. 相关工作

对于情感分类，早期的方法是基于支持向量机或统计模型，而最近的方法采用了递归神经网络等深度学习的方法[7]。一个文本序列通过Word2Vec[8]等词嵌入(word embedding)模型转化为词向量(word vector)序列作为模型的输入，使特征具有语义信息，深度学习构建网络模型模拟人脑神经系统对文本进行逐步分析、特征抽取并且自动学习优化模型输出，以提高分类的准确性[9]。情感分类任务可以划分为两大类：(1)传统深度学习的方法，典型代表CNN、LSTM和RNN等；(2)图神经网络的方法，典型代表有GCN等。本次实验中采用了传统深度学习方法中的基于卷积神经网络方法Bi-LSTM进行情感分类。

6. 未来展望

本文的情感分类任务是建立在有标签的数据上的，而现实生活中，有标签数据的获取是十分高昂的且十分繁琐，因为标签来源于人工标记，而人工判断的标准是有误差损失的，随着技术的发展，信息量是指数型增长，当处于不同的领域时，数据集的标注需要重新进行，因此现在大多数学者探寻以无监督学习的方式来进行情感分类任务，这里的无监督是不用标记数据的意思。

此次对微博评论的情感分类是二分类的，现实生活中人类的情感不会这么非好即坏的，因此二分类情感是非常粗粒度的，但是二分类情感分类速度快可以更快度得到一个总体情感极性，在一些情况下还是十分适用的。多数学者会使用[1，5]来衡量情感由极差到极好的过渡，其中“1”表示极差，“5”表示极好。因此，在后期对微博评论的分析的问题是更加细粒度和无监督情况下进行的。

Bi-LSTM参数多，使得训练难度加大了很多。因此后期可以使用效果和

Bi-LSTM相当但参数相对较少的GRU来构建大训练量的模型。

7. 总结

在本次实验中，通过前期调研工作大概了解了情感分析任务从前期到现在常用方法的演变过程，由于时间有限，实验过程中所用到的模型有常规数据概率统计模型如SVM和MultinomiaNB等，同时也尝试使用了一个深度学习模型Bi-LSTM，实验结果表明Bi-LSTM的效果是最佳的，同时jieba特征提取和词袋特征提取可以保留原始文本的最大的信息量，但相对于jieba特征提取，词袋特征提取的方法，更加耗时，因此jieba获胜。

参考文献

[1] https://zhuanlan.zhihu.com/p/27293420 2020/12/17

[2] Chan H P, Chen W, King I. A Unified Dual-view Model for Review Summarization and Sentiment Classification with Inconsistency Loss[C/OL]//Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: Association for Computing Machinery, 2020: 1191–1200[2020–10–28]. https://doi.org/10.1145/3397271.3401039. DOI:10.1145/3397271.3401039.

[3] Hochreiter S , Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[4] https://zhuanlan.zhihu.com/p/32085405 2020/12/16

[5] https://www.imooc.com/article/23821 2020/12/20 2020/12/21

[6] https://blog.csdn.net/u013385925/article/details/80385873 2020/12/28

[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Computation 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9. 8.1735

[8] NGUYEN H T,NGUYEN L M.Effective attention networks for aspect-level sentiment classification[C]//2018 10th International Conference on Knowledge and Systems Engineering (KSE).Ho Chi Minh City:Institude of Electrical and Electronics Engineers,2018:25-30

[9] 李胜旺,杨艺,许云峰,张妍.文本方面级情感分类方法综述[J].河北科技大学学报,2020,41(06):518-527.