微博文本挖掘中模型评估方法比较与分析*

于斐钥　李陶深**

广西大学计算机与电子信息学院，南宁，530004

摘　要　微博是现下国内互联网最大交流平台之一，它所产生的庞大数据背后蕴藏着巨大的商业价值和社会价值。人们试图对庞大的微博数据进行各种微博语料分析，以获取所需的字段或者有用信息。本文编码实现逻辑回归、朴素贝叶斯、支持向量机、xgboost、KNN 的分类器，实现对微博数据中的文本数据进行分类；采用准确率（ACC）、精确率（PPV）、回召率（SEV）、f1-score（F1）等若干个不同指标，在微博语料集下对不同分类器分类结果进行评估，并通过实验对比，对分类器的分类性能进行对比分析。

关键字　微博文本，XGboost，支持向量机，随机森林，朴素贝叶斯，逻辑回归，KNN

Comparison and Analysis of Model Evaluation Methods in Microblog Text Mining

Yu Feiyao　Li Taoshen

School of Computer, Electronics & Information Guangxi University
Nanning 530004,China　tshli@gxu.edu.cn

Abstract—Microblog is one of the largest communication platforms of the Internet in China. The huge data generated by microblog contains huge commercial value and social value. People try to analyze the huge microblog data to obtain the required fields or useful information. This paper encodes and implements classifiers such as logistic regression, naive Bayes, support vector machine(SVM), xgboost and KNN to classify text data in microblog data. And Then, using several different indicators such as accuracy (ACC), precision (PPV), recall (SEV) and F1 score (F1), the classification results of different classifiers are evaluated under the microblog corpus set. Finally, through the experimental comparison, the classification performance of the classifier is compared and analyzed.

Key words—Microblog text, XGboostt; Support vector machine(SVM), Random forest,naive Bayes,Logistic regression, KNN

1 引　言

微博是国内互联网最大社交平台之一，用户通过使用微博而产生的庞大数据因为不同内容针对不同的领域都有着不可忽视的影响，其背后蕴藏着巨大的商业价值和社会价值[1]。目前，人们试图对庞大的微博数据进行各种微博语料分析，以获取所需的字段或者有用信息。例如，在广告业，人们通过对微博广告数据进行统计分析，可以提高工作效率，快速寻找商品数据，有助于帮助商家提高销售效率。

文本数据挖掘技术是人们分析文本数据和处理文本数据的有效手段，已经广泛应用于各领域。针对微博数据的分析和挖掘主要涉及 NLP 处理相关技术，自然语言处理（Natural Language Processing，NLP）技术的主要内容之一就是情感分析。

情感分析（Sentimentanalysis）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[2]。比如某家电制造公司想了解自己生产的空调或者冰箱受客户用户喜爱程度，就可以从微博上用户对该公司的家电的性能的评论入手，通过识别商品评论的信息、判断客户的褒贬态度等，判断用户对该公司家电冰箱家电空调的评价是积极的还是消极的还是中性的评价。

情感分析作为互联网网页挖掘中新兴的一个领域，不同方向的研究者对其不同方面的研究也越来越多。文献[3]将微博的情感分析作为研究社交网络舆情的一项关键技术，提出了基于融合显性和隐性特征的中文微博情感分析方法。文献[2]构建了一个适合微博文本分析的情感词典，提出了微博语句文本的情感计算方法。文献[4]提出一种基于双语词典的微博多类情感分析方法，提高了分类效果与准确率。文献[5]提出了一个基于卷积神经网络的混合模型（CNN-SVM），

*基金项目：本文得到广西科技计划项目（桂科 AD20297125）和广西高等教育本科教学改革工程项目一般项目(2020JGA116 资助。

**通讯作者：李陶深，教授，tshli@gxu.edu.cn

可用于对微博事件的评论进行分类。文献[6]对航天微博进行情感分析，设计了一种基于 SVM 算法的能判断航天相关微博情感正极性和负极性的方法。文献[7]提出一种基于图的情感基准词选择方法，将情感词应用于挖掘短句情感特征，可更好地把握微博整句情感。

近年来，人们把机器学习方法应用于微博情感分析，取得了一些研究成果[8][9][10][11][12]。人们也利用逻辑回归（Logistic Regression，LR）、朴素贝叶斯(Naive Bayes，NB)、支持向量机（Support Vector Machines，SVM）、K 最近邻（K-Nearest Neighbor，KNN）等方法，实现对微博语句文本的分类[13][14][15][16]。

本文基于文献[1]提出的微博文本数据分析和挖掘方法，编码实现与优化了逻辑回归（LR）、朴素贝叶斯(NB)、支持向量机（SVM）、K 最近邻（KNN）、随机森林等方法的分类器，实现对微博数据中的文本数据进行情感分析和挖掘。最后采用准确率、精确度、回召率、f1-score 等若干个不同指标，对分类结果进行评估，交叉验证评估这几种算法的性能。

2 分类器及其实现

2．1 分类器简介

（1） 朴素贝叶斯（NB）

贝叶斯方法是一种研究不确定性的推理方法。不确定性常用贝叶斯概率表示，它是一种主观概率。经典概率代表事件客观存在，而贝叶斯概率则是随个人的主观认识的变化而变化[17]。投掷硬币可能出现的正反面两种情形，经典概率代表硬币正面朝上的概率，这是一个客观存在；而贝叶斯概率则指个人相信硬币会正面朝上的程度。例如，一个期权交易者认为“此时对 A 股做空能成功”的概率是 0.8，这个概率是交易者根据经验而成的个人判断。由于贝叶斯概率是主观判断，所以经常受到个人判断的多种因素影响而变化。

在众多的贝叶斯分类器算法中，朴素贝叶斯分类模型是最早的一种，其算法逻辑简单，朴素贝叶斯分类模型结构简单，其运算速度比同类算法快得多，分类所需时间短，并且在大多数情况下分类准确率较高，因此在实践中得到了广泛的应用。分类器有一个简单的假设:基于类条件独立属性的假设，即在给定的类状态条件下，属性是相互独立的[17]。

朴素贝叶斯是基于统计的简单但使用良好的分类器。在文本挖掘中，根据是否存在某些功能做出决策。这意味着根据训练数据将某个类的概率分配给每个要素。计算所有概率后，可以根据测试集中功能的存在做出决定。

朴素贝叶斯方法的优点有：分类效率稳定；对数据集较大的情况时用朴素贝叶斯方法可以比较快；对 NaN 不敏感，算法也比较简单，常用于文本分类；对较小量的数据集表现很好，能处理多分类任务。其缺点主要有：需计算先验概率；对输入数据的表达形式很敏感；分类决策存在错误率；对样本属性有关联的情况其分类效果不太好。

（2）支持向量机（SVM）

SVM 算法是基于向量维数理论和统计学习理论的结构风险最小化原理发展起来的、专门针对有限样本进行预测的一种新的机器学习方法。它根据有限的样本信息，在模型的复杂度和学习能力之间寻求最佳的权衡，以获得最佳的泛化能力。目前，SVM 已经初步显示出比现有方法好得多的性能。

SVM 算法有高准确率，适合较少量数据集的分类。它可以解决大型特征空间问题，处理非线性特征的相互作用，泛化能力比较强，在维度大的文本分类中流行[17]。但是它的内存消耗大，运行和调参也会花费很长时间。当数据集量很多时效率不高，且对非线性问题没有合适核函数，所以高维映射解释力不强，只支持线性 2 分类。

（3） 逻辑回归(LR)

逻辑回归是一种通用线性回归，可用于通过构造回归函数实现分类或预测。逻辑回归模型是一个专注于二进制分类问题的分类器，它还可以处理多分类的问题。逻辑回归将任何输入值映射到[0,1]，并获取线性回归中的预测值。然后，将此值映射到 Sigmoid 函数，并使用预测值作为 x 轴变量，将 y 轴用作概率。

逻辑回归属于判别式模型，该模型有很多正则化的方法，而且不必像在用朴素贝叶斯那样担心特征是否相关。与决策树、SVM 相比，逻辑回归的优点有：实现简单，广泛的应用于工程问题；分类时计算量小，速度快，内存消耗小；观测样本概率很方便；多重共线性使用 L2 正则化来解决；计算代价不高，易于理解和实现。但是，逻辑回归的主要缺点有：当特征空间很大时，逻辑回归的性能不好；容易欠拟合，准确度不高；不能很好地处理大量多类特征或变量；只能处理线性分类问题，对于非线性特征需要转换再处理。

（4）K 最近邻（KNN）

KNN 分类算法最初由 Cover 和 Hart 于 1968 年提出，是最简单的机器学习算法之一。该算法根据待分类样本 X 和已经分类好的训练样本间的距离，选择与测试集样本距离最小的 K 个样本，最后以 X 的 K 个最近邻中的大多数样本所属的类别作为 X 的类别。因为以未分类样本为圆心画出以 k 为半径的圆，圆中包含分类样本即已经训练过的样本就是已经知道类别清晰的样本，根据圆内分类样本的大多数的那一部分的类别作为未知样本即测试集样本的类别。因此比其他方法更适合于类域交叉或重叠较多的待划分样本集。

（5）XGboost

XGBoost 算法(Extreme Gradient Boosing)是陈天棋于2014年开发的一种基于GBDT (Gradient Boosting决策树)的推广算法[18]。传统的 GBDT 在训练模型时只使用一阶导数的信息。XGBoost 使用二阶泰勒展开，同时使用一阶和二阶导数。它还添加了控制复杂性的术语，以避免过拟合问题。此外，在每次迭代中引用随机森林策略来支持数据采样[18]。随机森林训练的决策树是相互独立的，但是 XGBoost 通过纠正以前决策树的错误来构建新的树。

XGBoost 的核心思想是不断增加树的数量，通过不断进行特性拆分来形成一棵新的树。只要增加一棵树，就是反复拟合之前的残差并学习新函数 mJ (xi)的过程;每次迭代后，给叶子节点一个学习率，以减少每棵树的比例，削弱其影响，从而达到有足够空间进行未来学习的目的;当训练后得到树时，可以输入所需测试样本数据的特征，在所有的树上找到对应的叶子节点——分别表示对应的结果；然后，对所有树的结果进行综合，得到测试数据的预测结果。

XGBoost 模型在应用中具有突出的优势。该模型实现了目标函数的二阶泰勒展开，有利于梯度下降更快、更准确，具有更高的精度；无需选择损失函数即可进行叶片分裂优化计算，提高了模型的适应性；正则项的加入可以减少过拟合问题，增强其泛化能力。参考随机森林算法支持行、列采样，既可以减少过拟合的可能性，又可以简化计算要求。但是，在选择最优分割点的预排序过程中，XGBoost 的空间复杂度和内存消耗过高。当数据量较大时，耗时较长。

（6）随机森林（RF）

分类方法需要考虑精确度和过拟合，而随机森林（RF）的精确度比传统方法高还无需考虑过拟合，因此可以通过聚集多个模型的组合方法来提高预测精度。该类方法首先利用训练集数据构建一组基本的分类模型，通过对每个基分类模型的预测值投票或取平均值来决定最终预测值[17]。

RF 方法是一种统计学习理论，作为一种分类和回归的集成学习算法在不同的领域中分别取得了不错的效果。它采取 Bootstrap 抽样法，通过多轮抽样，生成 k 个数据集并构成含有 k 棵决策树的随机森林。随机森林通过其随机性使得其不易陷入过拟合并降低敏感数据对实验预测结果的影响，通过投票得出最终预测结果。

2．2 分类器的实现

本文是基于文献[1]提出的技术实现编码与优化，且在适应本文数据集的情况下做了相应的改动，在某些细节的实现方面有所不同。

文献[1]在收集文本数据的过程中，每条微博单词量控制在150个以内。数据集包含每个采访者平均5.5条微博，共1000个采访者（大学生），故总共大约5000条微博数据，语料比较少，属于较少量样本。此外，文献[1]中使用的微博文本数据集不是开源的，文本特征标记是由12位专家手动一条条根据主动性人格特征分为0到9标注。

本文的数据集是通过爬虫程序，从指定的浏览器微博首页上爬取12万条的微博文本数据。爬取得到的用户微博内容放在csv表格中，然后手工分辨用户发布的微博内容，一条条判断，将消极语句复制粘贴到neg60000.txt中，将积极语句复制粘贴到pos60000.txt中，最终处理的结果是：消极语句6万条左右，积极语句6万条左右。但是在实验中，对12万条微博文本数据集进行分类器分类时运行时间比较长，例如SVM网格调参时长超过2天。为了提高实验分析的速度，我们选用3万条语句作为本文的微博文本数据集，其中消极语句1.5万条左右，积极语句1.5万条左右。

研究表明，数据集数量不同，相应的最优分类器也不同。本文是利用SKlearn 工具包来编码实现和优化各个分类器的。SKlearn是一个由Python语言第三方提供的非常强力的机器学习库，包含了从数据预处理到训练模型的各个方面。它拥有可以用于监督和无监督学习的方法，使用它可以极大的节省编写代码的时间和代码量。下面介绍6种分类器的实现方法。

（1）朴素贝叶斯分类器

SKlearn 工具包提供了3种基于朴素贝叶斯的分类算法，分别为高斯朴素贝叶斯（GaussianNB）、多项式朴素贝叶斯（MultinomialNB）和伯努利朴素贝叶斯（BernoulliNB）。MultinomialNB用于多项分布数据，BernoulliNB用于多重伯努利分布，是朴素贝叶斯文本分类的两大经典算法。因为本文的数据集是稀疏矩阵，而GaussianNB方法适用于密集矩阵，所以本文不考虑选用GaussianNB分类器。在使用未降维的数据集分别训练MultinomialNB和BernoulliNB时，BernoulliNB模型的准确率高；在使用降维的数据集训练MultinomialNB和BernoulliNB时，MultinomialNB报错不适用降维数据集中出现负值的情况。故使用BernoulliNB分类算法建立分类器。

为了建立朴素贝叶斯模型，本文从sklearn.naive_bayes库中引入BernoulliNB算法，使用BernoulliNB的fit()方法，即BernoulliNB.fit(X_train,y_train)，通过训练训练集建立一个伯努利朴素贝叶斯分类器。其中，X表示输入，y表示输出。调用的方法如下：

```
be=BernoulliNB()
be.fit=(X_train. y_train)
be.score=(X_test,y_test)
```

（2）KNN分类器

KNN算法是根据K值作为距离来分类的。参数K值确定后，在待分类的样本点中找到已经分类的K个点，K中哪一类越多，则就将其分为哪一类。

为了在KNN算法中寻找最好的K，本文首先从sklearn.neighbors中导入类KNeighbors Classifier，使用range(1,11)和for循环，来循环建立当K为1到11时的KNN分类器；然后再通过fit()函数训练模型，并使用score()方法得出此K时的模型准确率。在不同的K值下通过循环可找到模型评分最高的模型的K值，即最好的K距离。实验中，通过比较得出最高的准确率为0.7460时的最优K距离为9。

因为KNN的训练时间较少，所以使用了降维的数据集和未降维的数据集进行算法评估。

（3）SVM分类器

在SKlearn 工具中，SVM分类器的参数为惩罚系数C和核函数类型kernel。C指的是分类器的惩罚项，默认值为1.0。C值越高，对误分类样本的惩罚越大，所以训练样本中的准确率越高，但是其泛化能力降低，即测试数据的分类精度降低。反之，如果C减少，训练样本中允许有一些误分类，允许容错，其泛化能力大大增强。在训练样本中有噪声的情况下，通常使用后者，训练样本集中的误分类样本用作噪声[19]。

kernel指SVM分类器采用的核函数类型，SKlearn工具中可供选择的类型包括：linear（线性核函数）；poly（多项式核函数）；rbf（高斯核核函数）；sigmod（sigmod核函数）；precomputed（矩阵核函数），默认值为rbf。

本文首先通过GridsearchCV(svc,parameters2, cv=5, scoring='accuracy')算法进行网格调参，选择最优模型参数，进行五折交叉验证。然后从sklearn库中导入SVM算法，使用svm.SVC()建立四个不同核的分类器；再使用SVD降维后用train_test_split方法默认划分的训练集训练模型，通过svm.SVC.predict()用训练好的模型对测试集进行预测。最后导入metrics包，通过metrics的classification_report()方法得到四个分类器的评价。

因为本文使用的数据集为3万条微博数据，是文献[1]的数据集的6倍，所以SVM运行的时间很长。为了不要耗费更多的时间，必须使用降维过的数据。，

（4）Xgboost分类器

采用网格调参算法GridSearchCV获取最佳参数。数据集使用的是train_test_split()方法划分的TF-IDF向量化的原始数据x_和y。具体的操作步骤如下：首先调用XGBClassifier()建立一个默认参数的xgboost分类器；其次，使用网格调参算法GridSearchCV，得到参数max_depth在数据中的最优参数结果。

经过研究与实验对比，XGboost的最佳参数为{'gamma': 0.48275862068965514, 'learning_rate': 0.1, 'max_depth': 3,'n_estimators': 150}，故根据最佳参数建模。因为网格调参和最佳参数建立模型用了较长时间（因本文数据集较大），故仅使用降维的数据集。

具体编码实现时，从xgboost.sklearn库中引入XGBClassifier算法，使用XGBClassifier(gamma= 0.48, learning_rate= 0.1,max_depth= 3, n_estimators= 150)算法建立一个xgboost分类器，再使用fit(X_train,y_train)拟合训练。最后在对分类结果进行评估。

（5）逻辑回归（LR）

本文使用LogisticRegression()模型的默认参数建立逻辑回归模型。从sklearn.linear_model中引入LogisticRegression算法，使用以下调用方式来建立LR分类器，并用训练集拟合训练分类器：

LogisticRegression().fit(X_train, y_train)；

（6）随机森林

使用随机森林，可以大大节省调参和建模所需要耗费的时间。本文使用RandomForestClassifier()模型的默认参数，具体操作如下：从sklearn. ensemble中引入RandomForestClassifier算法，用以下调用方式来建立随机森林分类器并用训练集训练分类器：

RandomForestClassifier().fit(X_train, y_train)；

3 模型评估方法

3．1 模型评估数据集与实验环境设置

如前所述，本文为模型评估准备的数据集是通过爬虫程序，从指定的浏览器微博首页上爬取 3 万条的微博文本数据。其中 1.5 万条左右的消极语句复制粘贴到 neg60000.txt 中，其标注 lab 为 0；1.5 万条左右的积极语句复制粘贴到 pos60000.txt 中，其 lab 为 1。

用于模型评估的微机实验环境的设置如下：

CPU：Intel(R)Core(TM) i7-6700HQ 2.60 GHz
内存：16GB
硬盘：2TB
显示卡： NVIDIA GeForce GTX960M
操作系统: Windows10 操作系统
编程工具包：Anaconda 默认内嵌的 python3.7 和 conda、SKlearn 工具包、jieba、xgboost.sklearn
编译器： Jupyter notebook

3．2 模型评估方法与评估指标

模型评价或评估方法由以下三类：

第一大类是普通的准确率评价指标。可在 sklearn 中调用 score 方法获得给出模型在测试集上的大致准确性。

第二大类是精确率与召回率等指标。可在 sklearn 中调用 classification_ report 方法获得指标。classification_report 方法中没有直接给出指标数值的，一般使用混淆矩阵手动计算。

第三大类是 ROC 曲线与 AUC 值，这两个指标主要作为不平衡分类的评价指标。

混淆矩阵是用来反映某一个分类模型的分类结果的，其中行代表的是真实的类，列代表的是模型预测的分类。具体模型如表 1 所示，其中，a_{ij} 表示真实类为类 i 被预测为类 j 的数目。

表 1 混淆矩阵模型表

	预测类1	预测类2	…	预测类n
真实类1	a_{11}	a_{12}	…	a_{1n}
真实类2	a_{21}	a_{22}	…	a_{2n}
…	…	…	…	…
真实类n	a_{n1}	a_{n2}	…	a_{nn}

假设 TP（true positive）表示样本的真实类别为正，最后的预测类别为正；FP（false positive）表示样本的真实类别为负，最后的预测类别为正；FN（false negative）表示样本的真实类别为正，最后的预测类别为负；TN（true negative）表示样本的真实类别为负，最后的预测类别为负。以下是分类方法的评估指标：

（1）真正率（True Positive Rate，TPR）。也称为灵敏度、recall 和 SEN，计算公式如下：

$$SEN = \frac{TP}{TP + FN} \tag{1}$$

（2）假正率（False positive Rate，FPR）。计算公式如下：

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

（3）准确率（Accuracy），也称为 ACC。ACC 计算的是整体的平均准确性。对于不平衡分类，准确率不是好的衡量指标，并不能说明分类器分类实际的准确性。因此对于不平衡分类来说，PPV、SEN 和 F1 才是更好的衡量指标。其计算公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

（4）召回率（recall），即 SEN。实际上，SEN 测量分类器对某一类别的预测结果的覆盖范围一般是指将所有类别相加得到整体覆盖范围。其计算公式如下：

$$SEN = \frac{TP}{TP + FN} \tag{4}$$

（5）真负率（特异度），即 SPE。SPE 为在分类器分类结果中预测为消极的结果中真实为消极的比例。其计算公式如下：

$$SPE = \frac{TN}{TN + FP} \tag{5}$$

（6）精确度（Precision），也通常被称为 PPV。PPV 衡量分类器对某个类别的预测结果的准确性，将所有类别求和平均后得到整体的准确性。其计算公式如下：

$$PPV = \frac{TP}{TP + FP} \tag{6}$$

（7） F-score，通常也称为 F1。F1 衡量分类器对某个类别的预测结果的准确性和覆盖率，是准确率和召回率的调和平均数。其计算公式如下：

$$F1 = \frac{2 \times P \times R}{P + R} \tag{7}$$

（8）ROC（Receiver Operating Characteristic）曲线。ROC 曲线以 FPR 为横坐标，TPR 为纵坐标的曲线。在二分类问题中，实例的值往往是连续值，通过设定一个阈值将实例分为正类或负类。不同的阈值对应不同的分类，同时一个阈值对应于一个 FPR 及 TPR，并对应于 ROC 空间中的一个点 P，当阈值连续变化时，P 点也随即移动，最终绘成 ROC 曲线。ROC 曲线越贴近图的左上角，表示分类结果越准确。

（9）ROC 曲线下的面积（Area Under ROC Curve，AUC）。AUC 是指 ROC 曲线与横坐标轴所围区域的面积。通常取 0.5 到 1.0 之间的值，其值越大代表分类模型的性能越好。

3．2 模型评估与分析

为了验证数据准确率，对于文本数据需要对数据集进行划分。假设大写 X 表示输入，小写 y 表示输出。sklearn 机器学习库的 train_test_split 可以很方便进行划分数据集，对数据集进行快速打乱（分为训练集和测试集），默认 25%做为测试集，75%为训练集。因为 k 折越大越耗时间，本文使用了 KNN、SVM、XGboost、LR、随机森林、NB 等六个分类器，采用默认的交叉验证方式，只划分了一次数据集。原始默认采用的是 train_test_split 方法，使得数据划分具有偶然性。调用代码如下：

```
X_train,X_test,y_train,y_test=train_test_split(X,y)
```

因为本文的数据集较大，所以使用了降维处理数据集，并对在时间允许的范围内使用未降维数据集的 KNN、NB、LR 模型。进行降维与否对模型准确率等性能指标的比较。因为降维可以减小分类器建模和网格调参的时间，所以根据本文实验建立的 6 个分类模

型在分类过程所需的时间来看，数据集较大时，KNN、LR 和 NB 也可以使用未降维数据集进行训练和预测，即使用 TF-IDF（Term Frequency-Inverse Document Frequency）向量化的结果（代码表示中的 x_）。这里的 TF-IDF 是一种用于信息检索与数据挖掘的加权技术，用来评估一个字词对于语料库中的一份文件的重要程度[17]。而 SVM、XGboost、随机森林必须降维，否则一个分类器训练需要几天时间。LR 和 NB 算法不适合训练测试降维数据集，这会使模型准确率等指标变低分别为 0.79、0.66，尤其是 LR 和 NB 在分类未降维数据集的时候模型分类评估指标更高分别为 0.81、0.79。KNN 算法相比于分类未降维数据集更适合分类降维数据集，这会使 KNN 模型调参、训练测试时间变少且模型性能更优。KNN 模型分类未降维数据准确率为 0.60，模型分类降维数据准确率为 0.75。

对于已经训练好的模型分类测试集，使用以下不同的方法来得到模型分类的性能和准确率等指标。

①使用 score(X_test, y_test)方法，用测试集测试分类器算法，得出未降维的模型评分（score）或者说模型的准确率，降维的模型评分（score）或者说模型的准确率，以此分析各个分类器对未降维的模型和降维模型的适用性。

②使用 predict_proba(X_test)方法，对测试集合进行预测。再通过从 sklearn.metrics 导入得到 ROC、AUC 的方法 roc_curve()和 auc()。从 roc_curve(y_test, predictions)函数得到 FPR、TPR 的值，然后使用 matplotlib 包，以横轴和纵轴分别为 FPR、TPR 的值作图，来对模型进行评估。

从实验结果上看，在分类样本为中量数据集（3 万条）且考虑时间开销的情况下，NB 和 LR 分类模型可以使用未降维数据，KNN、SVM、XGboost、RF 分类模型最好使用降维数据，否则运行时间过长。

从实验结果还可以看到，在分类样本为中量数据集（3 万条微博数据）且考虑分类器准确率的情况下，NB 和 LR 分类模型使用未降维数据模型各指标更高；KNN 分类模型使用降维数据模型效果更好。SVM、XGboost、RF 因时间有限，未能使用未降维数据分类。即使使用降维数据分类，SVM、XGboost、RF 的耗时也很长。

为了使结果更具合理和说服力，本文全部使用降维数据分类，利用实验结果数据进行比对。

表 2 给出了朴素贝叶斯、SVM、KNN、XGboost、逻辑回归、随机森林等 6 个分类器的准确率评估结果，从表中可以看出，准确率最高的是随机森林分类器，其准确度为 0.80，其次是逻辑回归分类器和 XGboost 分类器，精确度最低的是贝叶斯分类器，准确度近为 0.66。而对于 ROC 曲线下的面积（AUC），XGboost、SVM 和逻辑回归分类器最高（0.88），最低的是贝叶斯（0.72）。所以在本文的数据集下，性能最好的分类器是 XGboost 分类器，因为性能根据 AUC 值判断，在 AUC 最高为 0.88 的三个算法中 XGboost 准确率最高为 0.79288；第二阶梯是 SVM 和逻辑回归分类器和随机森林，再次的是 KNN 分类器，性能最差的是朴素贝叶斯分类器。

表 2 6 种分类器的 score 评分（ACC）和 AUC 对比

分类器	准确率(ACC)	AUC
贝叶斯	0.665671	0.72
SVM	0.783172	0.88
KNN	0.746	0.82
XGboost	0.79288	0.88
逻辑回归	0.789644	0.88
随机森林	0.794374	0.87

对于 lab 为 1 的积极词，6 个分类器模型的不同性能指标如下表 3 所示。对于 lab 为 0 的消极词，6 个分类器模型的不同性能指标如下表 4 所示。

表 3 6 种分类器的对于积极词的性能指标 PPV、SEN、F1 对比

分类器	PPV	SEN	F1
贝叶斯	0.66	0.58	0.62
SVM	0.89	0.61	0.72
KNN	0.79	0.63	0.70
XGboost	0.83	0.70	0.76
逻辑回归	0.87	0.64	0.74
随机森林	0.82	0.72	0.77

从表 3 和表 4 可以看出，朴素贝叶斯分类器对于 lab 为 0 的消极词表现较好，分类性能 PPV、SEN、F1 指标值较高。因为 F1 指标综合了准确性与覆盖面，所以在结果不能用 PPV、SEN 单独衡量的时候，以 F1 值指标比较。如果以 F1 值指标作为比较标准，可以看出 SVM、KNN、XGboost、LR、RF 分类器对于 lab 为 0 的消极词的表现比朴素贝叶斯分类器更好，分类性能 F1 指标更高。故在本文数据集下，六种分类器（NB、SVM、KNN、XGboost、LR、RF）对于消极词的分类结果更准确、覆盖面更广、表现更好。综合消极词与积极词的整体预测指标 PPV、SEN、F1 如表 5 所示。

表 4 6 种分类器对于消极词的性能指标 PPV、SEN、F1 对比

分类器	PPV	SEN	F1
贝叶斯	0.67	0.74	0.71
SVM	0.73	0.93	0.82
KNN	0.72	0.85	0.78
XGboost	0.77	0.88	0.82
逻辑回归	0.75	0.92	0.82
随机森林	0.78	0.87	0.82

通过以上6个分类器之间的比较,可以看出SVM、随机森林、逻辑回归和XGboost分配器在所有指标上均表现出良好的性能,KNN和贝叶斯算法的性能不理想，这意味着它们可能不适合本论文数据集下的文本分类。这一实验结论与文献[1]的实验分析结论不太一样。

表 5 6种分类器的整体性能指标PPV、SEN、F1对比

分类器	PPV	SEN	F1
贝叶斯	0.665	0.66	0.71
SVM	0.81	0.77	0.82
KNN	0.755	0.74	0.78
XGboost	0.80	0.79	0.82
逻辑回归	0.81	0.78	0.82
随机森林	0.80	0.795	0.82

文献[1]的实验结论是：KNN、Xgboost不适合文本分类，SVM、LR表现稳定，而朴素贝叶斯分类器在AUC、PPV性能评估指标上表现最佳，但是在其它指标上表现不理想。本文的编码是在文献[1]的基础上实现的，并做了一定的优化处理。本文的实验分析结论与文献[1]的其结论不一样的原因有以下几点：

（1）文献[1]在收集文本数据的过程中，每条微博单词量控制在150个以内。本文采集数据时，只要是出现在微博网页上采集，单词量比文献[1]采集的文本数据的单词量要多得多。

表 6 6种分类器使用降维数据分类后得到的混淆矩阵

分类器	TN	TP	FN	FP	SPE	SEN
贝叶斯	3165	2183	1567	1119	0.7388	0.5821
SVM	4008	2284	1455	287	0.9332	0.6109
KNN	3649	2344	1406	635	0.8518	0.6251
XGboost	3760	2610	1129	535	0.8754	0.6981
逻辑回归	3933	2411	1328	362	0.9157	0.6448
随机森林	3733	2675	1055	571	0.8673	0.7172

（2）由于实验中的文本数据集有3万条微博数据，属于中量即样本，数据集的大小与SVM、Xgboost、随机森林分类器更兼容。而文献[1]的数据集只有大约5000条微博数据，属于较少量样本，使得朴素贝叶斯分类器表现最优。这说明了数据集数量不同，相应的最优分类器也不同。根据实验中6种分类器预测后得到的混淆矩阵，手动计算SPE和SEN，得到表6的结果。从表6可以看出，SPE始终高于SEN，表明此6种分类器更有能力识别具有低主动性人格的微博用户。显然，本文研究结论与文献[1]的研究结论基本一致。

4 结束语

本文编码实现了基于逻辑回归（LR）、朴素贝叶斯（NB）、支持向量机（SVM）、KNN、XGboost、随机森林方法的分类器，可用于对预处理后的微博文本数据集进行分类。采用准确率（ACC）、精确率（PPV）、回召率（SEV）、f1-score（F1）、PPV、SEN等若干个不同指标，在自建的微博网页数据集上对6个分类器的分类结果进行评估和对比分析，交叉验证评估这6种分类器的性能。通过实验对比，可以分析在以微博文本为数据集上，不同分类器在分析人们情感方面具有不同的适用度和分类效果。

参 考 文 献

[1] P. Wang, Y. Yan, Y. Si, et al.Classification of Proactive Personality: Text Mining Based On Weibo Text and Short-answer Questions Text[J].IEEE Access, 2020,8: 97370 - 97382

[2] 陈国兰.基于情感词典与语义规则的微博情感分析[J].情报探索, 2016,(2):1-6

[3] 陈铁明，缪茹一，王小号.融合显性和隐性特征的中文微博情感分析[J].中文信息学报. 2016,30(4)：184-192

[4] 栗雨晴,礼欣,韩煦等.基于双语词典的微博多类情感分析方法[J]. 电子学报，2016,44(9):2068-2073

[5] 涂曼姝,张艳,颜永红.基于 CNN-SVM 和转发树的微博事件情感分析[J].情报工程. 2017,3(3)：77-85

[6] 徐建忠,朱俊,赵瑞.基于SVM算法的航天微博情感分析[J].信息安全研究. 2017,3(12)：1129-1133

[7] 刘续乐,何炎祥.基于多特征的微博情感分析研究[J].计算机工程, 2017,43(12):160-164,172

[8] 颜扬君.基于深度学习的微博热点话题文本情感分析研究[D].南昌：南昌大学，2020

[9] 秦欣.基于深度学习的微博短文本情感分析技术研究[D].西安：西安建筑科技大学，2020

[10] 金志刚,胡博宏,张瑞.融合情感特征的深度学习微博情感分析[J].南开大学学报(自然科学版).2020,53(5):77 - 81,86

[11] 陈珂,叶颖雅,马乙平.用于微博情感分析的深度学习网络模型[J].计算机与数字工程. 2020,48(7):1674-1681

[12] 胡梦雅,樊重俊,朱玥.基于机器学习的微博评论情感分析[J].信息与电脑(理论版). 2020,32(12):71-73

[13] 申莹,刘春阳,赵永翼.基于SVM算法的微博评论数据情感分析[J].数字通信世界. 2020,(1):111,117

[14] 余鹏,田杰.基于卷积神经网络的多维特征微博文本情感分析[J].计算机与数字工程. 2020,48(9):2244-2247

[15] 王彬菁.基于朴素贝叶斯分类算法的微博文本的情感分析研究[J].中国新通信. 2019,21(08):114-115

[16] 冯军军,王海沛,贺晓春.基于 Logistic 回归模型的微博情感分析研究[J].计算机与数字工程. 2018,46(9): 1824- 1829，1843

[17] 王振武.大数据挖掘与应用[M].北京：清华大学出版社,2017

[18] 吕子夷. 基于机器学习算法的股指期货价格预测与比较研究[D].杭州：浙江大学,2020.

[19] 马梦曦. 基于弹幕文本挖掘的情感极性分析研究[D].武汉：武汉理工大学,2019.