**qq群标签分类器**

**一、数据**

1.train\_data.csv：训练数据4000

2.test\_data.csv：测试数据20462

**二、构建分类器的流程**

1.数据预处理

（1）词向量：基于“百度百科”中文词向量来构造。共300维

（2）词性标注：利用正态分布随机初始化。共50维

（3）连接词向量和词性标注：共350维

（4）在训练集中，利用5-cross\_validation划分训练集和验证集

|  |  |
| --- | --- |
| **Dataset** | **Number** |
| train dataset | 3201 |
| validation dataset | 799 |

2.分类器：lightgbm

|  |  |  |
| --- | --- | --- |
| **Hyper-Parameters** | **Description** | **Value** |
| max\_depth | 树的深度 | 10 |
| learning\_rate | 学习率 | 0.05 |
| n\_estimators | 评估器的数量 | 600 |
| objective | 目标函数 | binary |
| reg\_lambda | L2正则 | 0.5 |

**三、分类器效果（在验证集中评估）**

1.针对label 1(该标签需要去除的)的预测：

|  |  |  |  |
| --- | --- | --- | --- |
| **5 fold cross validation** | **Precision** | **Recall** | **F1-score** |
| 0 | 0.791667 | 0.500000 | 0.612903 |
| 1 | **0.875000** | 0.429825 | 0.576471 |
| 2 | 0.798611 | 0.504386 | 0.618280 |
| 3 | 0.823129 | 0.530702 | 0.645333 |
| 4 | 0.828947 | **0.555066** | **0.664908** |
| Mean | 0.823471 | 0.503996 | 0.623579 |

**四、预测**

1.根据构建的分类器，预测test\_data中的label不为1的数据。共预测**16896**条数据。

2.把预测值从大到小进行排列。

3.得到数据test\_pre\_label0.csv

|  |  |  |
| --- | --- | --- |
| **数量** | **描述** | **错误率** |
| 100个 | 100个中，有5个不应该删除 | 5% |

**qq群标签分类器2**

**一、数据**

1. tag\_origin.csv：最原始的标签

2. tags.50.filt.removeall.v6.del\_sig\_key.txt：删除了单个词和去掉了业务关键词的列表

3. 标注部分.xlsx：分为positive和negative数据

4. tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt：大部分是positive数据

5. tags.50.filt.removeall.v4.alltaginfo.txt：含有特征的数据

|  |  |
| --- | --- |
| **文件名** | **数量** |
| tag\_origin.csv | 24462 |
| tags.50.filt.removeall.v6.del\_sig\_key.txt | 15261 |
| 标注部分.xlsx（label\_positive.csv和label\_negative.csv） | Positive: 2977 Negative: 269 |
| tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt | Positive: 6752 |
| tags.50.filt.removeall.v4.alltaginfo.txt | 19758 |

**二、预处理方法**

1. “tag\_origin.csv”和“tags.50.filt.removeall.v6.del\_sig\_key.txt”做差集，得到大部分都为negative的词语，记为“**tags\_negative.txt**”。

2. “tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt”和“tags\_negative.txt”合并

1）“tag.both+top6k.final.review.raw.del\_sig\_ns\_key”和“tags\_negative.txt”做交集，如果出现共同的数据，则统一标为negative，“tag.both.txt”中剩下的数据都为positive，形成数据“**tag.both.remain.txt**”

3. “标注部分.xlsx”和“tag.both.remain.txt”合并 ，形成训练正例数据“**label\_positive.txt**”

1）“tag.both.remain.txt”和“标注部分.xlsx”中的negative做差集，即“tag.both.remain.txt”中剩下的数据则为positive，形成数据“**tag.both.remain\_pos.txt**”

2) “标注部分.xlsx”中的positive和“tag.both.remain\_pos.txt”做并集，得到positive的数据“**label\_positive.txt**”

4. “标注部分.xlsx”中的negative、“label\_positive”和“tags\_negative.txt”合并，形成训练负例数据“**label\_negative.txt**”

1）“tags\_negative.txt”和正例数据“label\_positive.txt”做差集，出现共同的数据标为positive, “tags\_negative.txt”中剩下的数据则为negative，形成数据“**tags\_negative\_remain.txt**”

2）“标注部分.xlsx”中的negative和“tags\_negative\_remain.txt”做并集，得到negative的数据“**label\_negative.txt**”

5.“label\_positive.txt和“label\_negative.txt”为训练数据，从“tags.50.filt.removeall.v4.alltaginfo.txt“除去训练数据，得到预测数据“test\_data.txt”。其中训练数据不在“tags.50.filt.removeall.v4.alltaginfo.txt“中的记为“featureless\_train\_data.txt”

|  |  |  |
| --- | --- | --- |
| **文件名** | **数量** | **描述** |
| **步骤1** | | |
| tags\_negative.txt | Negative: 9186 |  |
| **步骤2** | | |
| tag.both.remain.txt | Positive: 6752 | 比tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt去掉了0 |
| **步骤3** | | |
| tag.both.remain\_pos.txt | Positive: 6752 | 比tag.both.remain.txt去掉了0 |
| **label\_positive.txt** | **Positive: 7081** |  |
| **步骤4** | | |
| tags\_negative\_remain.txt | Negative: 9008 | 比tags\_negative.txt去掉了178 |
| **label\_negative.txt** | **Negative: 9008** |  |
| **步骤5** | | |
| test\_data.txt | 8368 |  |
| train\_data.txt | 11387  Positive: 7074  Negative: 4313 | label\_positive.txt + label\_negative.txt，并除去了featureless\_train\_data.txt |
| featureless\_train\_data.txt | 4702  Positive: 7  Negative: 4695 | 在train\_data.txt中没有特征可以用的数据 |

**qq群标签分类器3**

**一、数据**

1. tag\_origin.csv：最原始的标签

2. tags.50.filt.removeall.v6.del\_sig\_key.txt：删除了单个词和去掉了业务关键词的列表

3. 标注部分.xlsx：分为positive和negative数据

4. tags.50.filt.removeall.v4.alltaginfo.txt：含有特征的数据

|  |  |
| --- | --- |
| **文件名** | **数量** |
| tag\_origin.csv | 24462 |
| tags.50.filt.removeall.v6.del\_sig\_key.txt | 15261 |
| 标注部分.xlsx（label\_positive.csv和label\_negative.csv） | Positive: 2977 Negative: 269 |
| tags.50.filt.removeall.v4.alltaginfo.txt | 19758 |

**二、预处理方法**

1. “tag\_origin.csv”和“tags.50.filt.removeall.v6.del\_sig\_key.txt”做差集，得到大部分都为negative的词语，记为“**tags\_negative.txt**”。

2. “标注部分.xlsx”和“tags\_negative.txt”合并，形成训练正例数据“**label\_positive.txt**”和训练负例数据“**label\_negative.txt**”

1）“tags\_negative.txt”与“标注部分.xlsx”中的positive做差集，出现共同的数据标为positive, “tags\_negative.txt”中剩下的数据则为negative，形成数据“**tags\_negative\_remain.txt**”

2）“tags\_negative\_remain.txt”和“标注部分.xlsx”中的negative做并集，形成训练负例数据“**label\_negative.txt**”

3）“标注部分.xlsx”的positive形成训练正例数据“**label\_positive.txt**”

3.“label\_positive.txt和“label\_negative.txt”为训练数据，从“tags.50.filt.removeall.v4.alltaginfo.txt“除去训练数据，得到预测数据“test\_data.txt”。其中训练数据不在“tags.50.filt.removeall.v4.alltaginfo.txt“中的记为“featureless\_train\_data.txt”

4，“tags.50.filt.removeall.v6.del\_sig\_key.txt”和“test\_data.txt”做交集，得到新的训练数据“**test\_data2.txt**”

|  |  |  |
| --- | --- | --- |
| **文件名** | **数量** | **描述** |
| **步骤1** | | |
| tags\_negative.txt | Negative: 9186 |  |
| **步骤2** | | |
| tags\_negative\_remain.txt | Negative: 9008 | 比tags\_negative.txt去掉了178 |
| **label\_negative.txt** | **Negative: 9008** |  |
| **label\_positive.txt** | **Positive: 2975** | 2个是重复的 |
| **步骤3** | | |
| test\_data.txt | 12467 |  |
| test\_data2.txt | 12457 |  |
| train\_data.txt | 7288  Positive: 2975  Negative: 4313 | label\_positive.txt + label\_negative.txt，并除去了featureless\_train\_data.txt |
| featureless\_train\_data.txt | 4695  Positive: 0  Negative: 4695 | 在train\_data.txt中没有特征可以用的数据 |

**三、构建分类器的流程**

1.数据特征构造

（1）词向量：基于“百度百科”中文词向量来构造。共300维

（2）词性标注：利用正态分布随机初始化。共50维

（3）连接词向量和词性标注：共350维

（4）添加新的特征：6维+350维

2.切分训练集和验证集

在训练集中，利用5-cross\_validation划分训练集和验证集

|  |  |
| --- | --- |
| **Dataset** | **Number** |
| train dataset | 5831 |
| validation dataset | 1457 |

**三、分类器效果（在验证集中评估）**

1.针对label 1(该标签需要去除的)的预测：

|  |  |  |  |
| --- | --- | --- | --- |
| **5 fold cross validation** | **Precision** | **Recall** | **F1-score** |
| 0 | 0.9283196239717979 | 0.9154113557358053 | 0.9218203033838972 |
| 1 | 0.931899641577061 | 0.9038238702201622 | 0.9176470588235295 |
| 2 | 0.9457831325301205 | 0.9096176129779838 | 0.9273479031305375 |
| 3 | 0.9412484700122399 | 0.8921113689095128 | 0.9160214413341274 |
| 4 | 0.9556354916067147 | 0.9245939675174014 | 0.9398584905660379 |
| Mean | 0.940577 | 0.909112 | 0.924539 |

**四、预测**

1.根据构建的分类器，预测test\_data2中的数据。共预测**12457**条数据。

2.把预测值从大到小进行排列。

3.得到数据test\_pre.csv

4.随机抽样100条数据

|  |  |  |
| --- | --- | --- |
| **Confuse matrix** | **Predict 0** | **Predict 1** |
| **True 0** | 71 | 13 |
| **True 1** | 1 | 15 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Label** | **Number** | **Precision** | **Recall** | **F1-score** |
| 0 | 84 | 0.9861111111111112 | 0.8452380952380952 | 0.9102564102564102 |
| 1 | 16 | 0.535714 | 0.937500 | 0.681818 |

**qq群标签分类器4**

**一、分类器预测**

1.特征组合构造

|  |  |  |
| --- | --- | --- |
| **组合** | **分类器** | **特征** |
| 组合1 | SVM（RBF） | One-hot词性(41)+6维特征：47 |
| 组合2 | SVM（RBF） | 6维特征：6 |
| 组合3 | SVM（RBF） | 词向量+one-hot（词性）+6维特征：347 |
| 组合4 | Lightgbm | One-hot词性(41)+6维特征：47 |
| 组合5 | Lightgbm | 6维特征：6 |
| 组合6 | Lightgbm | 词向量+one-hot（词性）+6维特征：347 |
| 组合7 | SVM（线性核） | 6维特征：6 |

2.验证集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **组合** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合1 | 1 | 0.962675 | 0.663809 | 0.785764 |
| 组合2 | 1 | 0.988878 | 0.575474 | 0.727478 |
| 组合3 | 1 | 0.941446 | 0.708096 | 0.808235 |
| 组合4 | 1 | 0.935037 | 0.897055 | 0.915635 |
| 组合5 | 1 | 0.932519 | 0.890330 | 0.910925 |
| 组合6 | 1 | 0.938617 | 0.898911 | **0.918302** |
| 组合7 | 1 | 0.555616 | 0.787240 | 0.603088 |

3.100 sample预测

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **组合** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合1 | 0 | 0.835052 | 0.835052 | 0.895028 |
| 1 | 0 | 0 | 0 |
| 组合2 | 0 | 0.836735 | 0.976190 | 0.901099 |
| 1 | 0 | 0 | 0 |
| 组合3 | 0 | 0.836957 | 0.916667 | 0.875000 |
| 1 | 0.125000 | 0.062500 | 0.083333 |
| 组合4 | 0 | 0.973684 | 0.902439 | 0.936709 |
| 1 | **0.666667** | 0.888889 | **0.761905** |
| 组合5 | 0 | 0.986301 | 0.878049 | 0.929032 |
| 1 | 0.629630 | **0.944444** | 0.755556 |
| 组合6 | 0 | 0.948718 | 0.902439 | 0.925000 |
| 1 | 0.636364 | 0.777778 | 0.700000 |
| 组合7 | 0 | 0.820000 | 1.0 | 0.901099 |
| 1 | 0 | 0 | 0 |

4.数据预测

（1）tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt：删除了单个词、地名和去掉了业务关键词

（2）tags.50.filt.removeall.v6.del\_sig\_key.txt：删除了单个词和去掉了业务关键词的列表

|  |  |  |
| --- | --- | --- |
| 文件名 | Number | 得到特征后的number |
| tag.both+top6k.final.review.raw.del\_sig\_ns\_key.txt | 6752 | 6745 |
| tags.50.filt.removeall.v6.del\_sig\_key.txt | 15261 | 15254 |

5.观察组合6和组合3的验证集

（1）分类器单独输出的结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **组合** | **数量** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合6 | 1457 | 1 | 0.968474 | **0.890951** | **0.928097** |
| 组合3 | 1457 | 1 | **0.995842** | 0.555684 | 0.713328 |

（2）组合6和组合3共同进行预测：

|  |  |  |
| --- | --- | --- |
| **Confuse matrix** | **Predict 0** | **Predict 1** |
| **True 0** | 594 | 1 |
| **True 1** | 91 | 771 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** |
| 0 | 0.867153 | 0.998319 | 0.928125 |
| 1 | 0.998705 | 0.894432 | 0.943696 |

（3）各自不同的预测case

针对label 1，组合6辅助预测正确的：292

组合3辅助预测正确的：3

6.观察组合5和组合6的验证集

（1）分类器单独输出的结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **组合** | **数量** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合5 | 1457 | 1 | 0.961929 | 0.879350 | 0.918788 |
| 组合6 | 1457 | 1 | 0.968474 | 0.890951 | 0.928097 |

（2）组合5和组合6共同进行预测：

|  |  |  |
| --- | --- | --- |
| **Confuse matrix** | **Predict 0** | **Predict 1** |
| **True 0** | 579 | 16 |
| **True 1** | 80 | 782 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Label** | **Precision** | **Recall** | **F1-score** |
| 0 | 0.878604 | 0.973109 | 0.923445 |
| 1 | 0.979950 | 0.907193 | 0.942169 |

（3）各自不同的预测case

针对label 1，组合5辅助预测正确的：23

组合6辅助预测正确的：38

7.融合

（1）把两个组合的输出进行平均，大于0.6则标为1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **组合** | **数量** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合5 | 1457 | 1 | 0.961929 | 0.879350 | 0.918788 |
| 组合6 | 1457 | 1 | 0.968474 | 0.890951 | 0.928097 |
| 融合5和6 | 1457 | 1 | 0.975703 | 0.885151 | **0.928224** |

（2）100 sample：**lightgbm\_classifier\_add\_feature\_merge.py**

1）组合3和组合6进行融合

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **组合** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合3 | 0 | 0.836957 | 0.916667 | 0.875000 |
| 1 | 0.125000 | 0.062500 | 0.083333 |
| 组合6 | 0 | 0.948718 | 0.902439 | 0.925000 |
| 1 | 0.636364 | 0.777778 | 0.700000 |
| 融合 | 0 | 0.913580 | 0.902439 | 0.907975 |
| 1 | 0.592593 | **0.888889** | **0.711111** |

2）组合5和组合6进行融合

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **组合** | **Label** | **Precision** | **Recall** | **F1-score** |
| 组合5 | 0 | 0.986301 | 0.878049 | 0.929032 |
| 1 | 0.629630 | 0.944444 | **0.755556** |
| 组合6 | 0 | 0.948718 | 0.902439 | 0.925000 |
| 1 | 0.636364 | 0.777778 | 0.700000 |
| 融合 | 0 | 0.973333 | 0.890244 | 0.929936 |
| 1 | 0.640000 | 0.888889 | 0.744186 |

**qq群标签分类器分析**

**一、不同维度分析**

1.各个特征对测试集的影响

2.根据最后结果进行排序，给出前面的一些分析

3.观察train data的数据

**二、各个特征对测试集的影响**

1.抽样400个，并人工打上标记。生成文件名为：

**tags.50.filt.removeall.v6.del\_sig\_key.sample400.csv**

2.观察6个维度的特征对test data 的影响：针对的是label 1的预测

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **特征** | **阈值** | **precision** | **recall** | **f1\_score** |
| All feature | 0.7 | 0.460751 | 0.870968 | 0.602679 |
| - no newtag2cout | 0.7 | **0.552000** | 0.445161 | 0.492857 |
| - no newtitleexttag2count | 0.7 | 0.464164 | 0.877419 | 0.607143 |
| - no newdescexttag2count | 0.7 | 0.438202 | 0.754839 | 0.554502 |
| - no classifer | 0.7 | 0.422053 | 0.716129 | 0.531100 |
| - no entropy | 0.7 | 0.473310 | 0.858065 | 0.610092 |
| - no entropyv2 | 0.7 | 0.455479 | 0.858065 | 0.595078 |

3.组合不同的特征进行测试

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **特征** | **阈值** | **precision** | **recall** | **f1\_score** |
| newtag2cout + entropy | 0.7 | 0.373272 | 0.522581 | 0.435484 |
| newtag2cout + classifer | 0.7 | 0.414097 | 0.606452 | 0.492147 |

通过观察6个特征的重要性，得到每个特征的重要程度：

|  |  |
| --- | --- |
| **特征** | **重要程度** |
| newtag2cout | 8362 |
| newtitleexttag2count | 7067 |
| newdescexttag2count | 7325 |
| classifer | 7446 |
| entropy | 6916 |
| entropyv2 | 6418 |

**三、数据**

1.存在文件夹“analyse\_data”

**四、确定排序的阈值在哪个范围比较可信**

1.每个阈值区间中抽样100

|  |  |  |
| --- | --- | --- |
| **阈值** | **文件名** | **准确率** |
| 大于0.9 | tags.50.filt.removeall.v6.del\_sig\_key.test\_pre.100\_0.9.csv | 95% |
| 0.85-0.9 | tags.50.filt.removeall.v6.del\_sig\_key.test\_pre.100\_0.85\_0.9.csv | 79% |
| 0.8-0.85 | tags.50.filt.removeall.v6.del\_sig\_key.test\_pre.100\_0.8\_0.85.csv | 74% |

**五、观察train data的数据**

1.本身positive和negative数据存在近似的字样：

（1）单身、

（2）女汉纸、美女

（3）兼职

2.特征缺失值比较多

（1）train\_data数量为**7288**，缺失值的数量

|  |  |
| --- | --- |
| **特征** | **非空的数量** |
| newtag2cout | 6628 |
| newtitleexttag2count | 6040 |
| newdescexttag2count | 6104 |
| classifer | 5694 |
| entropy | 6040 |
| entropyv2 | 6040 |

（2）用算法进行预测填充: ./analyse\_data/train\_data\_no\_missing.csv

（3）利用得到的填充结果作为train data进行预测

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **特征** | **阈值** | **precision** | **recall** | **f1\_score** |
| All feature + word embedding | 0.7 | 0.487500 | 0.754839 | 0.592405 |
| All feature | 0.7 | 0.466912 | 0.819355 | 0.594848 |
| no newtag2cout+ word embedding | 0.7 | **0.606061** | 0.387097 | 0.472441 |
| no newtag2cout | 0.7 | 0.592233 | 0.393548 | 0.472868 |

（4）通过填充缺失值，可以提升**5%**左右的准确率。

**六、结论**

1.6个特征对分类器的重要性基本相同，也就是说明并没有出现比较强的特征可以指导分类器进行预测。

2.怀疑train data标记不是大部分正确。

3.在预测值为0.85以上的有较大置信度。

4.通过在train data上填充缺失值，提升5%左右的准确率。