# **Running results of different models on Sentiment 140**

## **Naive Bayes Model**

## Sentiment 140 Training set - Testing set

80,000 data is randomly selected for training data(setting random.seed in the code to ensure the 80,000 data extracted is the same every time) and the remaining 20,000 data is for testing.

Accuracy：0.74925

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8378 (true negatives) | 1634 (false positives) |
| Positive(real) | 3381 (false negatives) | 6607 (true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.74115

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8350 (true negatives) | 1662(false positives) |
| Positive(real) | 3515(false negatives) | 6473(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.7274

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8442(true negatives) | 1570(false positives) |
| Positive(real) | 3882(false negatives) | 6106(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.71675

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8323(true negatives) | 1689(false positives) |
| Positive(real) | 3976(false negatives) | 6012(true positives) |

The first 5,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.6987

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8224(true negatives) | 1788(false positives) |
| Positive(real) | 4238(false negatives) | 5750(true positives) |

### Sentiment 140 pure training set for testing

The first 5,000 data of the above 80,000 data is extracted for training and testing

Accuracy:0.9082

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 2416(true negatives) | 86(false positives) |
| Positive(real) | 373(false negatives) | 2125(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and testing

Accuracy:0.8895

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 4729(true negatives) | 243(false positives) |
| Positive(real) | 862(false negatives) | 4166(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and testing

Accuracy:0.8707

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 9410(true negatives) | 600(false positives) |
| Positive(real) | 1986(false negatives) | 8004(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and testing

Accuracy:0.856875

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 18302(true negatives) | 1561(false positives) |
| Positive(real) | 4164(false negatives) | 15973(true positives) |

The whole 80,000 data for training and testing

Accuracy：0.8395375

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 36249(true negatives) | 3589(false positives) |
| Positive(real) | 9248(false negatives) | 30914(true positives) |

1. **Vader Model（Based on rules, neutral sentiment judgments）**

Vader: 20,000 data for testing, same as 20,000 test set of naive bayes.

Accuracy:0.51845

|  |  |  |  |
| --- | --- | --- | --- |
|  | Negative(pred) | Neutral(predict) | Positive(pred) |
| Nagative(real) | 4227 | 2636 | 3149 |
| Neutral(real) | 0 | 0 | 0 |
| Positive(real) | 1003 | 2843 | 6142 |

## **Support Vector Machine**

main hyper parameter：C=1,kernel=’rbf’,random\_state=10

* Sentiment 140 Training set - Testing set

The first 5,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.6708

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6443(true negatives) | 3569(false positives) |
| Positive(real) | 3015(false negatives) | 6973(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.6758

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6468(true negatives) | 3544(false positives) |
| Positive(real) | 2940(false negatives) | 7048(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.6838

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6663(true negatives) | 3349(false positives) |
| Positive(real) | 2975(false negatives) | 7013(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.68675

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6588(true negatives) | 3424(false positives) |
| Positive(real) | 2841(false negatives) | 7147(true positives) |

The whole 80,000 data is for training and the same 20,000 data is for testing

Accuracy: 0.6937

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6602(true negatives) | 3410(false positives) |
| Positive(real) | 2716(false negatives) | 7272(true positives) |

### Sentiment 140 pure training set for testing

The first 5,000 data of the above 80,000 data is extracted for training and testing

Accuracy:0.8204

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 1967(true negatives) | 535(false positives) |
| Positive(real) | 363(false negatives) | 2135(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and testing

Accuracy: 0.8143

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 3858(true negatives) | 1114(false positives) |
| Positive(real) | 743(false negatives) | 4285(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and testing

Accuracy: 0.80745

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7788(true negatives) | 2222(false positives) |
| Positive(real) | 1629(false negatives) | 8361(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and testing

Accuracy: 0.801025

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 15229(true negatives) | 4634(false positives) |
| Positive(real) | 3325(false negatives) | 16812(true positives) |

The whole 80,000 data is extracted for training and testing

Accuracy: 0.794425

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 30258(true negatives) | 9580(false positives) |
| Positive(real) | 6866(false negatives) | 33296(true positives) |

## **Random Forest**

main hyper parameter：n-estimator=400, criterion='gini', max\_depth=10, random\_state=10

* Sentiment 140 Training set - Testing set

The first 5,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.6378

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6321(true negatives) | 3691(false positives) |
| Positive(real) | 3553(false negatives) | 6435(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.64455

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6073(true negatives) | 3939(false positives) |
| Positive(real) | 3170(false negatives) | 6818(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.64935

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6380(true negatives) | 3632(false positives) |
| Positive(real) | 3381(false negatives) | 6607(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.654

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6227(true negatives) | 3785(false positives) |
| Positive(real) | 3135(false negatives) | 6853(true positives) |

The whole 80,000 data is for training and the same 20,000 data is for testing

Accuracy: 0.65395

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6300(true negatives) | 3712(false positives) |
| Positive(real) | 3209(false negatives) | 6779(true positives) |

### Sentiment 140 pure training set for testing

The first 5,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.975

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 2384(true negatives) | 118(false positives) |
| Positive(real) | 7(false negatives) | 2491(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.9502

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 4514(true negatives) | 458(false positives) |
| Positive(real) | 40(false negatives) | 4988(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and testing

Accuracy： 0.9266

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 8890(true negatives) | 1120(false positives) |
| Positive(real) | 348(false negatives) | 9642(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.89025

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 16814(true negatives) | 3049(false positives) |
| Positive(real) | 1341(false negatives) | 18796(true positives) |

The whole 80,000 data is for training and testing

Accuracy：0.8374375

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 31901(true negatives) | 7937(false positives) |
| Positive(real) | 5068(false negatives) | 35094(true positives) |

## **Neural Network**

main hyper parameter：hidden layer:3, each layer has 256 hidden units; each hidden layer uses ‘relu’ as activation function. output layer: use sigmoid as activation function. opt='adam', epochs=10, batch\_size=32

* Sentiment 140 training set - testing set

The first 5,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.62485

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7647(true negatives) | 2365(false positives) |
| Positive(real) | 5138(false negatives) | 4850(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.64095

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6175(true negatives) | 3837(false positives) |
| Positive(real) | 3344(false negatives) | 6644(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.64755

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7546(true negatives) | 2466(false positives) |
| Positive(real) | 4583(false negatives) | 5405(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and the same 20,000 data is for testing

Accuracy: 0.66045

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 6922(true negatives) | 3090(false positives) |
| Positive(real) | 3701(false negatives) | 6287(true positives) |

The whole 80,000 data for training and the same 20,000 data is for testing

Accuracy: 0.66985

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7172(true negatives) | 2840(false positives) |
| Positive(real) | 3763(false negatives) | 6225(true positives) |

### Sentiment 140 pure training set for testing

The first 5,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.9172

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 2464(true negatives) | 38(false positives) |
| Positive(real) | 376(false negatives) | 2122(true positives) |

The first 10,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.9262

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 4502(true negatives) | 470(false positives) |
| Positive(real) | 268(false negatives) | 4760(true positives) |

The first 20,000 data of the above 80,000 data is extracted for training and testing

Accuracy：0.8888

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 9440(true negatives) | 570(false positives) |
| Positive(real) | 1654(false negatives) | 8336(true positives) |

The first 40,000 data of the above 80,000 data is extracted for training and testing

Accuracy： 0.87935

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 17849(true negatives) | 2014(false positives) |
| Positive(real) | 2812(false negatives) | 17325(true positives) |

The whole 80,000 data for training and testing

Accuracy： 0.8555125

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 35567(true negatives) | 4271(false positives) |
| Positive(real) | 7288(false negatives) | 32874(true positives) |

**Running results of different models in IMDb movie reviews**

10,000 pieces of data in total from 250 best movies and 100 worst movies respectively. 6 points and below 6 points are bad comments, 7 points and above 7 points are good comments. 7920 pieces of data were randomly grabbed as the test set (seed=10) and 1985 pieces of data were left as the test set In training dataset, there are 4714 positive reviews and 3206 negative reviews. In test dataset, there are 1186 positive reviews and 799 negative reviews

## **Naive Bayes**

### training set - testing set

The accuracy score is: 0.8473551637279597

The f1 score is: 0.8789452656811826

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 582(true negatives) | 217(false positives) |
| Positive(real) | 86(false negatives) | 1100(true positives) |

### pure training set for testing

The accuracy score is: 0.9142676767676767

The f1 score is: 0.9293370798210011

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 2776(true negatives) | 430(false positives) |
| Positive(real) | 249(false negatives) | 4465(true positives) |

### 

### **Support Vector Machine**

hyper parameter remains the same

### training set - testing set

The accuracy score is: 0.8806045340050378

The f1 score is: 0.9015371832156212

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 663(true negatives) | 136(false positives) |
| Positive(real) | 101(false negatives) | 1085(true positives) |

### pure training set for testing

The accuracy score is: 0.9361111111111111

The f1 score is: 0.9466244725738396

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 2927(true negatives) | 279(false positives) |
| Positive(real) | 227(false negatives) | 4487(true positives) |

1. **Random Forest**

max-depth was changed to 15，other hyper parameter remains the same（n-estimator=400, criterion='gini'，random\_state=10)

### training set - testing set

The accuracy score is: 0.8347607052896725

The f1 score is: 0.8703557312252964

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 556(true negatives) | 243(false positives) |
| Positive(real) | 85(false negatives) | 1101(true positives) |

### pure training set for testing

The accuracy score is: 0.9998737373737374

The f1 score is: 0.9998939442146568

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 3205(true negatives) | 1(false positives) |
| Positive(real) | 0(false negatives) | 4714(true positives) |

1. **Vader （Rule Based, no training set)**

* pure training set test

The accuracy score is: 0.7400503778337532

|  |  |  |  |
| --- | --- | --- | --- |
|  | Negative(pred) | Neutral(predict) | Positive(pred) |
| Nagative(real) | 456 | 2 | 341 |
| Neutral(real) | 0 | 0 | 0 |
| Positive(real) | 170 | 3 | 1013 |

1. **Neural Network**

hyper parameter remains the same

### training set - test set

The accuracy score is: 0.8705289672544081

The f1 score is: 0.8889848812095031

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 699(true negatives) | 100(false positives) |
| Positive(real) | 157(false negatives) | 1029(true positives) |

### pure training set test

The accuracy score is: 0.9848484848484849

The f1 score is: 0.9872367581365667

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 3159(true negatives) | 47(false positives) |
| Positive(real) | 73(false negatives) | 4641(true positives) |

**Using the training model from 8,000 comments of IMDB to predict 20,000 twitter (testing set)**

### **naive bayes**

The accuracy score is: 0.54635

The f1 score is: 0.5815229924818965

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 4623(true negatives) | 5389(false positives) |
| Positive(real) | 3684(false negatives) | 6304(true positives) |

### **support vector machine**

The accuracy score is: 0.582

The f1 score is: 0.5744247607411932

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 5998(true negatives) | 4014(false positives) |
| Positive(real) | 4346(false negatives) | 5642(true positives) |

### **random forest**

The accuracy score is: 0.547

The f1 score is: 0.4497084548104956

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7238(true negatives) | 2774(false positives) |
| Positive(real) | 6286(false negatives) | 3702(true positives) |

### 

### **neural network**

The accuracy score is: 0.5643

The f1 score is: 0.44882985452245416

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 7738(true negatives) | 2774(false positives) |
| Positive(real) | 6440(false negatives) | 3548(true positives) |

### **Using the training model from 80,000 Twitter comments to predict 1985 IMDB comments (testing set)**

### **naive bayes**

The accuracy score is: 0.7234256926952141

The f1 score is: 0.7658848614072497

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 538(true negatives) | 261(false positives) |
| Positive(real) | 288(false negatives) | 898(true positives) |

### **support vector machine**

The accuracy score is: 0.4100755667506297

The f1 score is: 0.04563977180114099

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 786(true negatives) | 13(false positives) |
| Positive(real) | 1158(false negatives) | 28(true positives) |

### **random forest**

The accuracy score is: 0.5717884130982368

The f1 score is: 0.5319383259911895

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 652(true negatives) | 147(false positives) |
| Positive(real) | 703(false negatives) | 483(true positives) |

### **nerual network**

The accuracy score is: 0.5768261964735516

The f1 score is: 0.5945945945945945

|  |  |  |
| --- | --- | --- |
|  | Negative(pred) | Positive(pred) |
| Negative(real) | 529(true negatives) | 270(false positives) |
| Positive(real) | 570(false negatives) | 616(true positives) |