Report

1. Through the full dataset, it’s obviously that the negative is much more than positive and neutral, false is more than true.
2. The image below, on the left is the metrics predicted by the DT model for 100 features.
   The image below, on the right is the metrics predicted by the DT model for 200 features.

	precision	recall	f1-score	support
negative	0.70	0.76	0.73	335
neutral	0.34	0.34	0.34	125
positive	0.08	0.03	0.04	40
accuracy			0.60	500
macro avg	0.38	0.38	0.37	500
weighted avg	0.56	0.60	0.58	500

	precision	recall	f1-score	support
negative	0.82	0.29	0.43	335
neutral	0.30	0.86	0.44	125
positive	0.09	0.05	0.06	40
accuracy			0.41	500
macro avg	0.40	0.40	0.31	500
weighted avg	0.63	0.41	0.40	500

As shown in the picture, the prediction accuracy of 100 features is significantly greater than 200 features(depends on accuracy). Because the number of features selected will affect the construction of the model.
About running time, the model of 200 features runs much longer than 100 features.

3. For test baseline predictors, I have intercepted the results of three standards models and compared them with VADER, as below.

BNB(Bernoulli Naive Bayes)

```
1 negative
2 negative
3 negative
4 negative
5 negative
6 neutral
7 negative
8 neutral
9 negative
10 negative
11 negative
12 negative
13 negative
14 negative
15 negative
16 negative
17 negative
18 neutral
19 negative
20 negative
21 negative
22 negative
23 neutral
24 negative
```

DT(Decision Trees)

```
1 negative
2 negative
3 neutral
4 positive
5 neutral
6 neutral
7 negative
8 neutral
9 negative
10 neutral
11 positive
12 negative
13 negative
14 negative
15 negative
16 negative
17 negative
18 neutral
19 negative
20 negative
21 negative
22 negative
23 neutral
```

MNB(Multinomial Naive Bayes)

```
1 negative
2 negative
3 negative
4 negative
5 negative
6 neutral
7 negative
8 neutral
9 negative
10 negative
11 neutral
12 negative
13 negative
14 negative
15 negative
16 negative
17 negative
18 neutral
19 negative
20 negative
21 negative
22 negative
23 neutral
24 negative
```

And then this is VADER:

```
1
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
2
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
3
{'neg': 0.251, 'neu': 0.749, 'pos': 0.0, 'compound': -0.6908}
4
{'neg': 0.0, 'neu': 0.863, 'pos': 0.137, 'compound': 0.4019}
5
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
6
{'neg': 0.074, 'neu': 0.698, 'pos': 0.228, 'compound': 0.5106}
7
{'neg': 0.0, 'neu': 0.867, 'pos': 0.133, 'compound': 0.3182}
8
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
9
{'neg': 0.0, 'neu': 0.89, 'pos': 0.11, 'compound': 0.3818}
10
{'neg': 0.158, 'neu': 0.633, 'pos': 0.209, 'compound': 0.2023}
11
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
12
{'neg': 0.0, 'neu': 0.885, 'pos': 0.115, 'compound': 0.296}
13
{'neg': 0.083, 'neu': 0.827, 'pos': 0.09, 'compound': 0.0516}
14
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
15
{'neg': 0.158, 'neu': 0.842, 'pos': 0.0, 'compound': -0.4588}
16
{'neg': 0.167, 'neu': 0.833, 'pos': 0.0, 'compound': -0.4588}
17
{'neg': 0.266, 'neu': 0.55, 'pos': 0.183, 'compound': -0.2023}
18
{'neg': 0.179, 'neu': 0.821, 'pos': 0.0, 'compound': -0.34}
19
{'neg': 0.0, 'neu': 0.741, 'pos': 0.259, 'compound': 0.5994}
20
{'neg': 0.18, 'neu': 0.82, 'pos': 0.0, 'compound': -0.5106}
21
{'neg': 0.165, 'neu': 0.506, 'pos': 0.329, 'compound': 0.7096}
22
{'neg': 0.128, 'neu': 0.872, 'pos': 0.0, 'compound': -0.4215}
23
{'neg': 0.0, 'neu': 0.849, 'pos': 0.151, 'compound': 0.6249}
```

According to VADER's correct rate calculation method(depends on compound). I found that the correct rate of VADER is lower than these models. I think the reason is VADER is a simple model and the correct rate calculated by this model is definitely not as good as the other three more complex models.

4. For test remove stop words,we can see the image below, on the left is the metrics predicted by three models and the image on the right is the metrics predicted by three models deleted stop words.

DT:

```
0.414
0.40416288064985423
0.414
0.3127943170122818
              precision    recall  f1-score   support

    negative       0.82      0.29      0.43       335
     neutral       0.30      0.86      0.44       125
    positive       0.09      0.05      0.06        40

    accuracy                           0.41       500
   macro avg       0.40      0.40      0.31       500
weighted avg       0.63      0.41      0.40       500
```

```
0.528
0.3112088265283504
0.528
0.3114174324700641
              precision    recall  f1-score   support

    negative       0.66      0.72      0.69       335
     neutral       0.21      0.18      0.19       125
    positive       0.06      0.05      0.06        40

    accuracy                           0.53       500
   macro avg       0.31      0.31      0.31       500
weighted avg       0.50      0.53      0.51       500
```

BNB:

```
0.596
0.37542021105557016
0.596
0.369976299049104
              precision    recall  f1-score   support

    negative       0.70      0.76      0.73       335
     neutral       0.34      0.34      0.34       125
    positive       0.08      0.03      0.04        40

    accuracy                           0.60       500
   macro avg       0.38      0.38      0.37       500
weighted avg       0.56      0.60      0.58       500
```

```
0.734
0.4835907335907336
0.734
0.4271756064463814
              precision    recall  f1-score   support

    negative       0.74      0.98      0.84       335
     neutral       0.71      0.32      0.44       125
    positive       0.00      0.00      0.00        40

    accuracy                           0.73       500
   macro avg       0.48      0.43      0.43       500
weighted avg       0.67      0.73      0.67       500
```

MNB:

```
0.74
0.7317261282778524
0.74
0.5168195112114925
              precision    recall  f1-score   support

    negative       0.77      0.93      0.84       335
     neutral       0.59      0.42      0.49       125
    positive       0.83      0.12      0.22        40

    accuracy                           0.74       500
   macro avg       0.73      0.49      0.52       500
weighted avg       0.73      0.74      0.71       500
```

```
0.746
0.7905682339825563
0.746
0.5335677804911628
              precision    recall  f1-score   support

    negative       0.79      0.92      0.85       335
     neutral       0.59      0.49      0.53       125
    positive       1.00      0.12      0.22        40

    accuracy                           0.75       500
   macro avg       0.79      0.51      0.53       500
weighted avg       0.75      0.75      0.72       500
```

Obviously, for the three models, deleting stop words increases the accuracy of the prediction to varying degrees. The reason is some words have no distinction, these useless words will affect the final predictions.

5. On the left side is the predict results and on the right side is the correct rate.

DT:

```
1 negative
4 negative
5 positive
6 negative
7 negative
8 negative
9 negative
10 negative
11 positive
12 positive
13 negative
14 negative
16 negative
18 negative
19 negative
20 negative
22 negative
23 negative
24 positive
```

```
0.8133333333333334
0.5296965119472673
0.8133333333333334
0.5307807807807808
              precision    recall  f1-score   support

    negative       0.90      0.89      0.89       335
    positive       0.16      0.17      0.17        40

    accuracy                           0.81       375
   macro avg       0.53      0.53      0.53       375
weighted avg       0.82      0.81      0.82       375
```

BNB:

```
1 negative
4 negative
5 negative
6 negative
7 negative
8 negative
9 negative
10 negative
11 negative
12 negative
13 negative
14 negative
16 negative
18 negative
19 negative
20 negative
22 negative
23 negative
24 negative
```

```
0.896
0.9478609625668449
0.8960000000000001
0.4968867178093502
```

	precision	recall	f1-score	support
negative	0.90	1.00	0.94	335
positive	1.00	0.03	0.05	40
accuracy			0.90	375
macro avg	0.95	0.51	0.50	375
weighted avg	0.91	0.90	0.85	375

MNB:

```
1 negative
4 negative
5 negative
6 negative
7 negative
8 negative
9 negative
10 negative
11 negative
12 negative
13 negative
14 negative
16 negative
18 negative
19 negative
20 negative
22 negative
23 negative
24 negative
```

```
0.9173333333333333
0.8830216744581385
0.9173333333333333
0.6853090062532146
```

	precision	recall	f1-score	support
negative	0.92	0.99	0.96	335
positive	0.85	0.28	0.42	40
accuracy			0.92	375
macro avg	0.88	0.63	0.69	375
weighted avg	0.91	0.92	0.90	375

VADER:

```
1
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
2
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
3
{'neg': 0.251, 'neu': 0.749, 'pos': 0.0, 'compound': -0.6908}
4
{'neg': 0.0, 'neu': 0.863, 'pos': 0.137, 'compound': 0.4019}
5
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
6
{'neg': 0.074, 'neu': 0.698, 'pos': 0.228, 'compound': 0.5106}
7
{'neg': 0.0, 'neu': 0.867, 'pos': 0.133, 'compound': 0.3182}
8
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
9
{'neg': 0.0, 'neu': 0.89, 'pos': 0.11, 'compound': 0.3818}
10
{'neg': 0.158, 'neu': 0.633, 'pos': 0.209, 'compound': 0.2023}
11
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
12
{'neg': 0.0, 'neu': 0.885, 'pos': 0.115, 'compound': 0.296}
13
{'neg': 0.083, 'neu': 0.827, 'pos': 0.09, 'compound': 0.0516}
```

Through the pictures we can see that after the neutral sentiment is removed, the accuracy of the prediction is greatly improved. The reason is neutral sentiment is sometimes difficult to judge, and may involve irony, implied meaning, etc.

6. About my best method for sentiment analysis and my best method for topic classification, i used MNB model and some methods to preprocessed data. I deleted all kinds of special letters and symbols that are not related to emotions.

```
0.9173333333333333
0.8830216744581385
0.9173333333333333
0.6853090062532146
              precision    recall  f1-score   support

    negative       0.92      0.99      0.96       335
    positive       0.85      0.28      0.42        40

    accuracy                           0.92       375
   macro avg       0.88      0.63      0.69       375
weighted avg       0.91      0.92      0.90       375
```