COMP9414 assignment2

Huiyao zuo

Z5196480

1.Give simple descriptive statistics showing the frequency distributions for the sentiment and topic classes across the full dataset. What do you notice about the distributions?

In the 2000 data , most of them are negative, then the neutral ,the positive is the smallest one which is only 7.65%. The top famous topic is 10003 economic management ,almost cover one fifth of all data . Two topic take over 10% part and some other close to 10%,the rest topic only cover little part of the data ,most less one is 0.23%(10009). Both the sentiment and topic distribute are not balance.

2.Vary the number of words from the vocabulary used as training features for the methods (e.g. the top N features for N = 100,200, etc.). Show metrics calculated on both the training set and the test set. Explain any difference in performance of the models between training and test set, and comment on metrics and runtimes in relation to the number of features.

the sentiment part
Set the N features = 200

	precision	recall	f1-score	support
negative	0.72	0.86	0.78	335
neutral	0.44	0.33	0.37	125
positive	0.33	0.05	0.09	40
accuracy			0.66	500
macro avg	0.50	0.41	0.42	500
weighted avg	0.62	0.66	0.63	500

0.014260053634643555

Set the N features = 150

```
              precision    recall  f1-score   support

    negative       0.72      0.86      0.78       335
     neutral       0.44      0.33      0.37       125
    positive       0.33      0.05      0.09        40

    accuracy                           0.66       500
   macro avg       0.50      0.41      0.42       500
weighted avg       0.62      0.66      0.63       500

0.014053106307983398
```

Set the N features = 120

```
              precision    recall  f1-score   support

    negative       0.72      0.86      0.78       335
     neutral       0.44      0.33      0.37       125
    positive       0.33      0.05      0.09        40

    accuracy                           0.66       500
   macro avg       0.50      0.41      0.42       500
weighted avg       0.62      0.66      0.63       500

0.008847236633300781
```

Set the N features = 100

```
              precision    recall  f1-score   support

    negative       0.72      0.86      0.78       335
     neutral       0.44      0.33      0.37       125
    positive       0.33      0.05      0.09        40

    accuracy                           0.66       500
   macro avg       0.50      0.41      0.42       500
weighted avg       0.62      0.66      0.63       500

0.011049985885620117
```

the topic part
Set the N features = 200

```
    accuracy                           0.31       500
   macro avg       0.15      0.16      0.15       500
weighted avg       0.26      0.31      0.27       500

0.03370404243469238
```

Set the N features = 150

```
    accuracy                           0.29       500
   macro avg       0.15      0.15      0.14       500
weighted avg       0.26      0.29      0.26       500

0.033368110656738 28
```

Set the N features = 120

```
    accuracy                           0.29       500
   macro avg       0.15      0.15      0.14       500
weighted avg       0.26      0.29      0.26       500

0.025829076766967773
```

Set the N features = 100

```
    accuracy                           0.27       500
   macro avg       0.15      0.14      0.14       500
weighted avg       0.25      0.27      0.24       500

0.024637937545776367
```

The results show that with the features number decrease , the sentiment accuracy have no obvious change , the learning time is fluctuated but in the overall trend ,the learning time is decrease with the features number decrease. However ,the accuracy of topic could change with the features number in the same trend ,as same as the change trend of learning time . The reason of this could be the topic has more class than the sentiment ,so that it would need more words to distinguish the difference between topics. An other reason is the sentiment classes is very unbalance which would lead most words are from one class and learner would very easy to find the major sentiment.

3.Evaluate the standard models with respect to baseline predictors (VADER for sen- timent analysis, majority class for topic classification). Comment on the performance of the baselines and of the methods relative to the baselines.

The accuracy of VADER is 0.2685
The majority class for sentiment is 0.647
The majority class for topic is 0.179

the sentiment part

DT_sentiment features = 200

	precision	recall	f1-score	support
negative	0.72	0.86	0.78	335
neutral	0.44	0.33	0.37	125
positive	0.33	0.05	0.09	40
accuracy			0.66	500
macro avg	0.50	0.41	0.42	500
weighted avg	0.62	0.66	0.63	500

0.014260053634643555

BNB_sentiment

	precision	recall	f1-score	support
negative	0.71	0.98	0.83	335
neutral	0.74	0.22	0.34	125
positive	0.00	0.00	0.00	40
accuracy			0.71	500
macro avg	0.48	0.40	0.39	500
weighted avg	0.66	0.71	0.64	500

MNB_sentiment

	precision	recall	f1-score	support
negative	0.79	0.89	0.84	335
neutral	0.55	0.51	0.53	125
positive	1.00	0.10	0.18	40
accuracy			0.73	500
macro avg	0.78	0.50	0.52	500
weighted avg	0.75	0.73	0.71	500

The topic part

DT_topic features = 200

accuracy			0.31	500
macro avg	0.15	0.16	0.15	500
weighted avg	0.26	0.31	0.27	500

BNB_topic

accuracy			0.18	500
macro avg	0.03	0.05	0.02	500
weighted avg	0.07	0.18	0.06	500

MNB_topic

accuracy			0.27	500
macro avg	0.16	0.11	0.11	500
weighted avg	0.29	0.27	0.22	500

All the methods have better predict accuracy than the baseline , the MNB learner have best accuracy of sentiment and the best accuracy of topic is DT learner. The BNB learner have better accuracy than the DT learner in the sentiment .however ,it have the smallest accuracy in distinguish topic.

4.Evaluate the effect that preprocessing the input features, in particular stop word removal and Porter stemming as implemented in NLTK, has on classifier performance, for the three methods for both sentiment and topic classification. Compare results with and without preprocessing on training and test sets and comment on any similarities and differences.

After apply the stop word removal and Porter stemming , the accuracy all grow ,because the words in sentence have significant reduce. For the DT learner , all accuracy have little rise that could caused by the feature words slightly changed in the top list .for the reason that some words with special mark can not be deal by the NLTK tools.for the MNB learner , there is a great grow in the topic

accuracy. Because the MNB learner would use all words to learn ,delete the useless words would significant improve the accuracy of the learner.

DT_sentiment features = 200

accuracy			0.69	500
macro avg	0.41	0.40	0.39	500
weighted avg	0.61	0.69	0.63	500

DT_topic features = 200

accuracy			0.34	500
macro avg	0.23	0.21	0.21	500
weighted avg	0.33	0.34	0.32	500

BNB_sentiment

accuracy			0.72	500
macro avg	0.44	0.43	0.43	500
weighted avg	0.64	0.72	0.67	500

BNB_topic

accuracy			0.20	500
macro avg	0.09	0.06	0.03	500
weighted avg	0.18	0.20	0.09	500

MNB_sentiment

accuracy			0.73	500
macro avg	0.60	0.52	0.53	500
weighted avg	0.71	0.73	0.71	500

MNB_topic

accuracy			0.40	500
macro avg	0.30	0.21	0.22	500
weighted avg	0.41	0.40	0.36	500

5.Sentiment classification of neutral tweets is notoriously difficult. Repeat the experiments of items 3, 2 and 4 for sentiment analysis with only the positive and negative tweets (i.e. removing neutral tweets from both training and test sets). Compare these results to the previous results. Is there any difference in the metrics for either of the classes (i.e. consider positive and negative classes individually)?
5.2
DT_sentiment features = 200

	precision	recall	f1-score	support
negative	0.91	0.98	0.94	448
positive	0.38	0.10	0.16	49
accuracy			0.90	497
macro avg	0.65	0.54	0.55	497
weighted avg	0.86	0.90	0.87	497

DT_sentiment features = 150

	precision	recall	f1-score	support
negative	0.91	0.98	0.94	448
positive	0.38	0.10	0.16	49
accuracy			0.90	497
macro avg	0.65	0.54	0.55	497
weighted avg	0.86	0.90	0.87	497

DT_sentiment features = 100

	precision	recall	f1-score	support
negative	0.91	0.98	0.94	448
positive	0.38	0.10	0.16	49
accuracy			0.90	497
macro avg	0.65	0.54	0.55	497
weighted avg	0.86	0.90	0.87	497

DT_topic features = 200

	precision	recall	f1-score	support
10000	0.36	0.43	0.39	65
10001	0.19	0.32	0.24	37
10002	0.51	0.50	0.51	36
10003	0.19	0.53	0.28	85
10004	0.00	0.00	0.00	2
10005	0.71	0.51	0.60	49
10006	0.03	0.02	0.02	46
10007	0.00	0.00	0.00	2
10008	0.00	0.00	0.00	52
10009	0.00	0.00	0.00	2
10010	0.00	0.00	0.00	13
10011	0.00	0.00	0.00	7
10012	0.00	0.00	0.00	7
10013	0.00	0.00	0.00	32
10014	0.00	0.00	0.00	2
10015	0.67	0.52	0.59	23
10016	0.00	0.00	0.00	12
10017	0.00	0.00	0.00	6
10018	0.00	0.00	0.00	7
10019	0.00	0.00	0.00	12
accuracy			0.28	497
macro avg	0.13	0.14	0.13	497
weighted avg	0.24	0.28	0.24	497

DT_topic features = 150

	precision	recall	f1-score	support
10000	0.36	0.43	0.39	65
10001	0.19	0.32	0.24	37
10002	0.51	0.50	0.51	36
10003	0.19	0.53	0.28	85
10004	0.00	0.00	0.00	2
10005	0.71	0.51	0.60	49
10006	0.03	0.02	0.02	46
10007	0.00	0.00	0.00	2
10008	0.00	0.00	0.00	52
10009	0.00	0.00	0.00	2
10010	0.00	0.00	0.00	13
10011	0.00	0.00	0.00	7
10012	0.00	0.00	0.00	7
10013	0.00	0.00	0.00	32
10014	0.00	0.00	0.00	2
10015	0.67	0.52	0.59	23
10016	0.00	0.00	0.00	12
10017	0.00	0.00	0.00	6
10018	0.00	0.00	0.00	7
10019	0.00	0.00	0.00	12
accuracy			0.28	497
macro avg	0.13	0.14	0.13	497
weighted avg	0.24	0.28	0.24	497

DT_topic features = 100

	precision	recall	f1-score	support
10000	0.36	0.43	0.39	65
10001	0.19	0.32	0.24	37
10002	0.51	0.50	0.51	36
10003	0.19	0.53	0.28	85
10004	0.00	0.00	0.00	2
10005	0.71	0.51	0.60	49
10006	0.03	0.02	0.02	46
10007	0.00	0.00	0.00	2
10008	0.00	0.00	0.00	52
10009	0.00	0.00	0.00	2
10010	0.00	0.00	0.00	13
10011	0.00	0.00	0.00	7
10012	0.00	0.00	0.00	7
10013	0.00	0.00	0.00	32
10014	0.00	0.00	0.00	2
10015	0.67	0.52	0.59	23
10016	0.00	0.00	0.00	12
10017	0.00	0.00	0.00	6
10018	0.00	0.00	0.00	7
10019	0.00	0.00	0.00	12
accuracy			0.28	497
macro avg	0.13	0.14	0.13	497
weighted avg	0.24	0.28	0.24	497

5.3
The majority class for sentiment is 0.894
The majority class for topic is 0.187

5.4
DT_sentiment features = 200

	precision	recall	f1-score	support
negative	0.93	0.96	0.95	448
positive	0.48	0.31	0.38	49
accuracy			0.90	497
macro avg	0.71	0.64	0.66	497
weighted avg	0.88	0.90	0.89	497

DT_topic features = 200

	precision	recall	f1-score	support
10000	0.35	0.42	0.38	65
10001	0.41	0.24	0.31	37
10002	0.44	0.56	0.49	36
10003	0.17	0.44	0.24	85
10004	0.00	0.00	0.00	2
10005	0.82	0.55	0.66	49
10006	0.14	0.11	0.12	46
10007	0.00	0.00	0.00	2
10008	0.82	0.35	0.49	52
10009	0.00	0.00	0.00	2
10010	0.50	0.54	0.52	13
10011	0.00	0.00	0.00	7
10012	0.00	0.00	0.00	7
10013	0.00	0.00	0.00	32
10014	0.00	0.00	0.00	2
10015	0.60	0.65	0.63	23
10016	0.00	0.00	0.00	12
10017	0.00	0.00	0.00	6
10018	0.00	0.00	0.00	7
10019	0.00	0.00	0.00	12
accuracy			0.33	497
macro avg	0.21	0.19	0.19	497
weighted avg	0.36	0.33	0.32	497

From the data, the accuracy is change in the same trend with normal , but after apply the NLTK tools, the precision in different topics have big change , even the accuracy is grow ,some topics' precision is lower. And all the accuracy is better than the baseline.

6.Describe your best method for sentiment analysis and your best method for topic classification. Give some experimental results showing how you arrived at your methods. Now provide a similar evaluation of your best methods in relation to the standard methods and the baselines considering all the above issues.

Since the MNB learner have the highest accuracy in sentiment and topic after apply the NLTK tools.so I choose the MNB learner as the method used in my model.and I find the Porter stemming

accuracy			0.36	500
macro avg	0.20	0.18	0.17	500
weighted avg	0.35	0.36	0.32	500

would change all words to lower case ,so I try not use the Porter stemming tool .but the accuracy is lower in sentiment
And topics

	precision	recall	f1-score	support
negative	0.82	0.83	0.82	335
neutral	0.52	0.62	0.56	125
positive	0.33	0.10	0.15	40
accuracy			0.72	500
macro avg	0.56	0.51	0.51	500
weighted avg	0.70	0.72	0.70	500

So I keep all the NLTK tools in the model.
Then I try to change the value of alpha to try the best value to get higher accuracy .
For sentiment ,when the value of alpha equal to 1.5 ,the accuracy will grow a little and keep the value over 2 then decrease.so I choose the value of alpha equal to 1.5.

accuracy			0.74	500
macro avg	0.65	0.51	0.51	500
weighted avg	0.73	0.74	0.72	500

For topics, when the value of alpha equal to 0.4 ,the accuracy will grow a little and keep the value over 0.8 then decrease.so I choose the value of alpha equal to 0.4.

accuracy			0.42	500
macro avg	0.36	0.28	0.29	500
weighted avg	0.42	0.42	0.41	500

In conclusion , this method have the highest accuracy than other standard methods and the baseline. Use the NLTK tools, add one ’s’ to the stop words list and change the value of alpha to get the model The difference in sentiment with other methods is very small ,like 0.01, and the largest gap is 0.05 with DT learner. For the topics part , the biggest difference to standard model is obvious , 0.22 to the BNB learner.