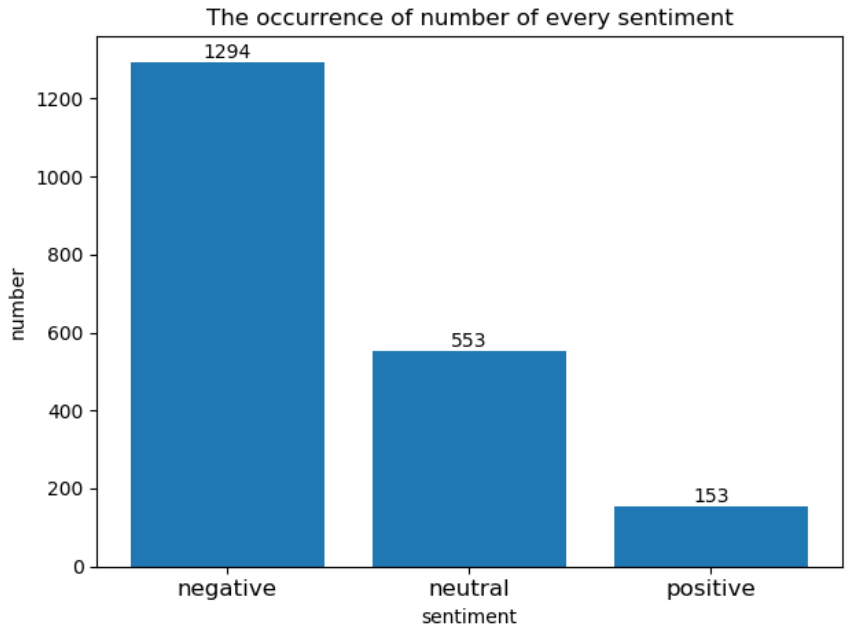
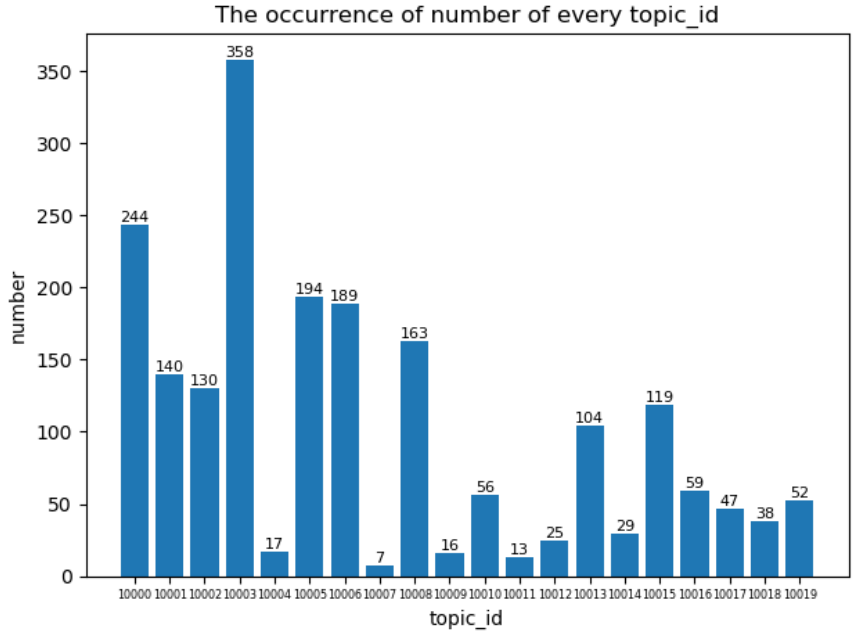
Name: Sun Zhenhang Student ID: z5231733

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?

   Answer:

   According to the dataset, we can get two bar charts about the distribution of sentiment and topic respectively. They are shown below:





   From the chart of sentiment, we can find that the “negative” class has the highest frequency, however the “positive” class has the lowest. With respect to the chart of topic, it is easy to see that class “10003” occurs most frequently and class “10007” hardly occurs. And most classes have the frequency that between 50 and 250.

2. Vary the number of words from the vocabulary used as training features for the standard methods (e.g. the top N words 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.

Answer:

One kind of table is showing metrics, whose p represents precision, r refers to recall. Another kind of table is about classification reports. In these both tables, the red numbers stand for the results of the test set as well as black figures are for the outcomes of the training set. And the other table represents the runtimes of training in relation to the number of features.

2.1 DT model (for sentiment):

Table 2-1-1 The metrics of DT model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.682/0.697	0.682/0.697	0.459/0.580	0.682/0.697	0.403/0.455	0.682/0.697	0.400/0.468

Table 2-1-2 The metrics of DT model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.688/0.699	0.688/0.699	0.470/0.575	0.688/0.699	0.420/0.466	0.688/0.699	0.416/0.479

Table 2-1-3 The metrics of DT model (top 300 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.688/0.699	0.688/0.699	0.470/0.575	0.688/0.699	0.420/0.466	0.688/0.699	0.416/0.479

And for the tables of metrics, when N = 400, 500, etc. (no more than 1000), the results are same as the outcome of N = 200.

Table 2-1-4 The classification report of DT model (top 100 words)

N=100	Precision	Recall	F1-score	Support
Negative	0.73/0.72	0.91/0.92	0.81/0.81	335/959
Neutral	0.45/0.59	0.27/0.34	0.38/0.43	125/428
Positive	0.20/0.42	0.03/0.10	0.04/0.16	40/113
Micro avg	0.68/0.70	0.68/0.70	0.68/0.70	500/1500
Macro avg	0.46/0.58	0.40/0.46	0.40/0.47	500/1500
Weighted avg	0.62/0.66	0.68/0.70	0.63/0.66	500/1500

Table 2-1-5 The classification report of DT model (top 200 words)

N=100	Precision	Recall	F1-score	Support
Negative	0.74/0.74	0.90/0.91	0.81/0.81	335/959
Neutral	0.47/0.56	0.34/0.39	0.39/0.46	125/428
Positive	0.20/0.42	0.03/0.10	0.04/0.16	40/113
Micro avg	0.69/0.70	0.69/0.70	0.69/0.70	500/1500
Macro avg	0.47/0.58	0.42/0.47	0.42/0.48	500/1500
Weighted avg	0.63/0.67	0.69/0.70	0.65/0.66	500/1500

Table 2-1-6 The classification report of DT model (top 300 words)

N=100	Precision	Recall	F1-score	Support
Negative	0.74/0.74	0.90/0.91	0.81/0.81	335/959
Neutral	0.47/0.56	0.34/0.39	0.39/0.46	125/428
Positive	0.20/0.42	0.03/0.10	0.04/0.16	40/113
Micro avg	0.69/0.70	0.69/0.70	0.69/0.70	500/1500
Macro avg	0.47/0.58	0.42/0.47	0.42/0.48	500/1500
Weighted avg	0.63/0.67	0.69/0.70	0.65/0.66	500/1500

Similarly, for the tables of classification reports, when N = 400, 500, etc. (no more than 1000), the results are same as the outcome of N = 200.

Comment on DT models for sentiment analysis:

From above tables, it is easy to get that the training set has higher metrics values than the test set, that is because the model is trained based on the training set.

Normally, with the increase of the number of features, the performance of the model will improve, however when the number of features becomes so large, it will cause overfitting problem, in this case, adding more features will hardly change the performance of the model.

2.2 DT model (for topics):

Table 2-2-1 The metrics of DT model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.272/0.350	0.272/0.350	0.159/0.187	0.272/0.350	0.144/0.183	0.272/0.350	0.143/0.177

Table 2-2-2 The metrics of DT model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.302/0.385	0.302/0.385	0.188/0.239	0.302/0.385	0.164/0.218	0.302/0.385	0.166/0.221

And for the tables of metrics, when N = 300, 400, etc. (no more than 1000), the results are same as the outcome of N = 200.

Table 2-2-3 The classification report of DT model (top 100 words)

N=100	Precision	Recall	F1-score	support
10000	0.23/0.35	0.36/0.51	0.28/0.42	56/188
10001	0.41/0.47	0.31/0.26	0.35/0.33	36/104
10002	0.43/0.43	0.48/0.51	0.45/0.46	31/99
10003	0.19/0.27	0.46/0.60	0.27/0.37	87/271
10004	0.00/0.00	0.00/0.00	0.00/0.00	2/15
10005	0.74/0.63	0.50/0.51	0.60/0.57	52/142
10006	0.14/0.18	0.14/0.12	0.14/0.15	44/145
10007	0.00/0.00	0.00/0.00	0.00/0.00	2/5
10008	0.19/0.23	0.07/0.09	0.10/0.13	46/117
10009	0.00/0.00	0.00/0.00	0.00/0.00	4/12
10010	0.00/0.00	0.00/0.00	0.00/0.00	11/45

10011	0.00/0.00	0.00/0.00	0.00/0.00	7/6
10012	0.00/0.00	0.00/0.00	0.00/0.00	4/21
10013	0.20/0.17	0.11/0.16	0.14/0.17	37/67
10014	0.00/0.00	0.00/0.00	0.00/0.00	6/23
10015	0.65/0.84	0.46/0.75	0.54/0.79	24/95
10016	0.00/0.00	0.00/0.00	0.00/0.00	14/45
10017	0.00/0.17	0.00/0.14	0.00/0.16	12/35
10018	0.00/0.00	0.00/0.00	0.00/0.00	10/28
10019	0.00/0.00	0.00/0.00	0.00/0.00	15/37
Micro avg	0.27/0.35	0.27/0.35	0.27/0.35	500/1500
Macro avg	0.16/0.19	0.14/0.18	0.14/0.18	500/1500
Weighted avg	0.27/0.31	0.27/0.35	0.25/0.31	500/1500

Table 2-2-4 The classification report of DT model (top 200 words)

N=200	Precision	Recall	F1-score	support
10000	0.29/0.36	0.41/0.54	0.34/0.43	56/188
10001	0.41/0.47	0.31/0.26	0.35/0.33	36/104
10002	0.48/0.49	0.42/0.45	0.45/0.47	31/99
10003	0.24/0.30	0.56/0.57	0.34/0.39	87/271
10004	0.00/0.00	0.00/0.00	0.00/0.00	2/15
10005	0.55/0.53	0.52/0.51	0.53/0.52	52/142
10006	0.15/0.30	0.16/0.26	0.16/0.27	44/145
10007	0.00/0.00	0.00/0.00	0.00/0.00	2/5
10008	0.20/0.42	0.09/0.29	0.12/0.34	46/117
10009	0.00/0.00	0.00/0.00	0.00/0.00	4/12
10010	0.60/0.39	0.27/0.27	0.37/0.32	11/45
10011	0.00/0.00	0.00/0.00	0.00/0.00	7/6
10012	0.00/0.00	0.00/0.00	0.00/0.00	4/21
10013	0.18/0.37	0.08/0.21	0.11/0.27	37/67
10014	0.00/0.00	0.00/0.00	0.00/0.00	6/23
10015	0.65/0.84	0.46/0.75	0.54/0.79	24/95
10016	0.00/0.00	0.00/0.00	0.00/0.00	14/45
10017	0.00/0.17	0.00/0.11	0.00/0.14	12/35
10018	0.00/0.17	0.00/0.14	0.00/0.15	10/28
10019	0.00/0.00	0.00/0.00	0.00/0.00	15/37
Micro avg	0.30/0.38	0.30/0.38	0.30/0.38	500/1500
Macro avg	0.19/0.24	0.16/0.22	0.17/0.22	500/1500
Weighted avg	0.28/0.36	0.30/0.38	0.27/0.36	500/1500

Similarly, for the tables of classification reports, when N = 300, 400, etc. (no more than 1000), the results are same as the outcome of N = 200.

Comment on DT models for topic analysis:

Most conclusions are same as the comment of 2.1. However, there exists some differences. One is that in table 2-2-3 and table 2-2-4, we can find that some classes' metrics values are 0, which means the model did not predict these classes based on

the given sets. This could be caused by the size of the training set, which is too small to build a high performance model to fit most cases.

2.3 BNB model (for sentiment):

Table 2-3-1 The metrics of BNB model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.726/0.721	0.726/0.721	0.579/0.631	0.726/0.721	0.508/0.547	0.726/0.721	0.528/0.573

Table 2-3-2 The metrics of BNB model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.726/0.759	0.726/0.759	0.625/0.704	0.726/0.759	0.545/0.649	0.726/0.759	0.568/0.672

Table 2-3-3 The metrics of BNB model (top 300 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.730/0.773	0.730/0.773	0.675/0.716	0.730/0.773	0.554/0.661	0.730/0.773	0.585/0.683

Table 2-3-4 The metrics of BNB model (top 400 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.722/0.789	0.722/0.789	0.629/0.751	0.722/0.789	0.549/0.675	0.722/0.789	0.573/0.704

Comment on BNB model for sentiment analysis:
And for the tables of metrics of BNB model, we can find that when the number of N gets larger, the training set has higher metrics values, however, the metrics values of the test set sometimes get higher sometimes lower. And we find that, when N equals to 800 (whose accuracy is 0.734), the test set has the highest metrics values.

2.4 BNB model (for topics):

Table 2-4-1 The metrics of BNB model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.264/0.401	0.264/0.401	0.173/0.391	0.264/0.401	0.149/0.266	0.264/0.401	0.153/0.294

Table 2-4-2 The metrics of BNB model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.330/0.507	0.330/0.507	0.213/0.472	0.330/0.507	0.183/0.326	0.330/0.507	0.186/0.353

Table 2-4-3 The metrics of BNB model (top 300 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.346/0.538	0.346/0.538	0.214/0.426	0.346/0.538	0.194/0.330	0.346/0.538	0.192/0.349

Table 2-4-4 The metrics of BNB model (top 400 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.362/0.576	0.362/0.576	0.212/0.518	0.362/0.576	0.191/0.346	0.362/0.576	0.190/0.367

Comment on BNB model for topic analysis:

And for the tables of metrics of BNB model, we can find that when the number of N gets larger, the training set has higher metrics values, however, the metrics values of the test set sometimes get higher sometimes lower. And we find that, when N equals to 400 (whose accuracy is 0.362), the test set has the highest metrics values.

2.5 MNB model (for sentiment):

Table 2-5-1 The metrics of MNB model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.732/0.723	0.732/0.723	0.612/0.550	0.732/0.723	0.502/0.655	0.732/0.723	0.529/0.581

Table 2-5-2 The metrics of MNB model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.734/0.761	0.734/0.761	0.665/0.710	0.734/0.761	0.538/0.637	0.734/0.761	0.567/0.665

Table 2-5-3 The metrics of MNB model (top 300 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.730/0.780	0.730/0.780	0.651/0.724	0.730/0.780	0.560/0.662	0.730/0.780	0.588/0.688

Table 2-5-4 The metrics of MNB model (top 400 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.740/0.797	0.740/0.797	0.717/0.766	0.740/0.797	0.589/0.695	0.740/0.797	0.627/0.724

Comment on MNB model for sentiment analysis:

And for the tables of metrics of MNB model, we can find that when the number of N gets larger, the training set has higher metrics values, however, the metrics values of the test set sometimes get higher sometimes lower. And we find that, when N equals to 400 (whose accuracy is 0.740), the test set has the highest metrics values.

2.6 MNB model (for topics):

Table 2-6-1 The metrics of MNB model (top 100 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.256/0.408	0.256/0.408	0.152/0.398	0.256/0.408	0.139/0.263	0.256/0.408	0.140/0.292

Table 2-6-2 The metrics of MNB model (top 200 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.324/0.524	0.324/0.524	0.215/0.522	0.324/0.524	0.191/0.387	0.324/0.524	0.196/0.422

Table 2-6-3 The metrics of MNB model (top 300 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.334/0.575	0.334/0.575	0.196/0.600	0.334/0.575	0.193/0.425	0.334/0.575	0.190/0.465

Table 2-6-4 The metrics of MNB model (top 400 words)

Accuracy	Micro-p	Macro-p	Micro-r	Macro-r	Micro-f1	Macro-f1
0.356/0.618	0.356/0.618	0.224/0.653	0.356/0.618	0.211/0.458	0.356/0.618	0.212/0.499

Comment on MNB model for topic analysis:

And for the tables of metrics of MNB model, we can find that when the number of N gets larger, the training set has higher metrics values, however, the metrics values of the test set sometimes get higher sometimes lower. And we find that, when N equals to 800 (whose accuracy is 0.384), the test set has the highest metrics values.

2.7 Runtimes:

Table 2-7-1 The runtimes of training of DT model for sentiment

N	100	200	300	400	500	600	700
Time(ms)	8.05	15.96	23.93	28.96	37.93	43.85	52.86

Table 2-7-2 The runtimes of training of BNB model for sentiment

N	100	200	300	400	500	600	700
Time(ms)	6.96	9.95	12.96	13.97	14.96	17.92	19.95

Table 2-7-3 The runtimes of training of MNB model for sentiment

N	100	200	300	400	500	600	700
Time(ms)	5.95	6.95	7.95	8.91	8.98	10.97	11.94

Table 2-7-4 The runtimes of training of DT model for topic

N	100	200	300	400	500	600	700
Time(ms)	9.91	22.96	30.92	39.86	48.91	60.84	69.85

Table 2-7-5 The runtimes of training of BNB model for topic

N	100	200	300	400	500	600	700
Time(ms)	4.94	7.99	9.98	11.93	13.96	14.96	16.90

Table 2-7-6 The runtimes of training of MNB model for topic

N	100	200	300	400	500	600	700
Time(ms)	4.10	5.92	6.98	7.97	7.98	9.90	10.91

Comment on runtimes:

According to all tables of runtimes, we can get a conclusion that more features cost more time to train. Besides, with the same N, MNB algorithm costs the least time and DT algorithm costs the largest time.

3. Evaluate the standard models with respect to baseline predictors (VADER for sentiment analysis, majority class for both classifiers). Comment on the performance of the baselines and of the methods relative to the baselines.

Answer:

For sentiment analysis, the majority class is "negative" which has 959 instances in the training set, and in the test set, the "negative" class has 335 instances. So the majority class baseline is 335/500 = 0.67. With respect to VADER baseline, we get the accuracy is 0.430. Compared with standard models, the results are shown below (the accuracy of standard models):

Table 3-1 The results of standard models, Majority class and VADER (aim for sentiment analysis)

DT	BNB	MNB	Majority class	VADER
0.688	0.716	0.738	0.670	0.430

For topic analysis, the majority class is "10003" which has 271 instances in the training set, and in the test set, the "10003" class has 87 instances. So the majority class baseline is 87/500 = 0.174. Compared with standard models, the results are shown below (the accuracy of standard models):

Table 3-2 The results of standard models and Majority class (aim for top analysis)

DT	BNB	MNB	Majority class
0.302	0.180	0.288	0.174

Comment:

From the table 3-1, we can see that the accuracy of VADER is the lowest which means it is difficult to anticipate the sentiment, and all the accuracy of standard models is above the majority class baseline. It is easy to see that MNB has the highest accuracy. And from the table 3-2, we can get a conclusion that DT is more suitable for topic analysis, and similarly, all the accuracy of standard models is above the majority class baseline.

4. Evaluate the effect that preprocessing the input features, in particular stop word removal plus Porter stemming as implemented in NLTK, has on classifier performance, for the three standard 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.

Answer:

In these both tables, the red numbers stand for the results of the test set as well as black figures are for the outcomes of the training set (here preprocessing means stop words removal and stemming).

Table 4-1 The metrics of DT model (top 200 words) of sentiment analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.688/0.699	0.470/0.575	0.420/0.466	0.416/0.479
With preprocessing	0.678/0.691	0.390/0.452	0.386/0.414	0.371/0.403

Table 4-2 The metrics of BNB model of sentiment analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.716/0.832	0.471/0.590	0.410/0.558	0.400/0.561
With preprocessing	0.712/0.835	0.436/0.579	0.429/0.563	0.421/0.562

Table 4-3 The metrics of MNB of sentiment analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.738/0.937	0.641/0.959	0.523/0.818	0.536/0.866
With preprocessing	0.744/0.925	0.591/0.945	0.546/0.809	0.552/0.857

From above tables, we can see that for some models (i.e. DT and BNB), adding stop words removal and stemming did not increase the metrics value for sentiment analysis. For MNB model, implementing preprocessing may increase the metrics value slightly.

Table 4-4 The metrics of DT model (top 200 words) of topic analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.302/0.385	0.188/0.239	0.164/0.218	0.166/0.221
With preprocessing	0.338/0.405	0.208/0.215	0.202/0.221	0.198/0.212

Table 4-5 The metrics of BNB model of topic analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.180/0.244	0.028/0.054	0.053/0.075	0.020/0.048
With preprocessing	0.198/0.285	0.036/0.172	0.061/0.094	0.031/0.071

Table 4-6 The metrics of MNB of topic analysis

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.288/0.719	0.173/0.799	0.125/0.453	0.124/0.505
With preprocessing	0.378/0.763	0.288/0.793	0.195/0.527	0.204/0.583

However, for tables of topic analysis, we find that using stop words removal and stemming can increase the metrics value.

5. Sentiment classification of neutral tweets is notoriously difficult. Repeat the experiments of items 2 (with N = 200), 3 and 4 for sentiment analysis with the standard models using 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)?

Answer:

In these both tables, the red numbers stand for the results of the test set as well as black figures are for the outcomes of the training set.

For item 2 (N=200):

Table 5-1 The classification report of DT model (have “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.74/0.74	0.90/0.91	0.81/0.81	335/959
Neutral	0.47/0.56	0.34/0.39	0.39/0.46	125/428
Positive	0.20/0.42	0.03/0.10	0.04/0.16	40/113
Micro avg	0.69/0.70	0.69/0.70	0.69/0.70	500/1500
Macro avg	0.47/0.58	0.42/0.47	0.42/0.48	500/1500
Weighted avg	0.63/0.67	0.69/0.70	0.65/0.66	500/1500

Table 5-2 The classification report of DT model (no “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.91/0.92	0.99/0.98	0.95/0.95	335/959
Positive	0.62/0.65	0.20/0.25	0.30/0.36	40/113
Micro avg	0.90/0.91	0.90/0.91	0.90/0.91	375/1072
Macro avg	0.76/0.78	0.59/0.62	0.62/0.65	375/1072
Weighted avg	0.88/0.89	0.90/0.91	0.88/0.89	375/1072

Table 5-3 The classification report of BNB model (have “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.79/0.81	0.86/0.86	0.82/0.83	335/959
Neutral	0.55/0.66	0.53/0.60	0.54/0.63	125/428
Positive	0.53/0.65	0.25/0.49	0.34/0.56	40/113
Micro avg	0.73/0.76	0.73/0.76	0.73/0.76	500/1500
Macro avg	0.62/0.70	0.54/0.65	0.57/0.67	500/1500
Weighted avg	0.71/0.75	0.73/0.76	0.71/0.75	500/1500

Table 5-4 The classification report of BNB model (no “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.93/0.95	0.98/0.97	0.95/0.96	335/959
Positive	0.65/0.68	0.38/0.55	0.48/0.61	40/113
Micro avg	0.91/0.93	0.91/0.93	0.91/0.93	375/1072
Macro avg	0.79/0.81	0.68/0.76	0.71/0.78	375/1072
Weighted avg	0.90/0.92	0.91/0.93	0.90/0.92	375/1072

Table 5-5 The classification report of MNB model (have “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.79/0.80	0.88/0.88	0.83/0.84	335/959
Neutral	0.56/0.66	0.51/0.58	0.54/0.62	125/428
Positive	0.64/0.66	0.23/0.45	0.33/0.54	40/113
Micro avg	0.73/0.76	0.73/0.76	0.73/0.76	500/1500
Macro avg	0.66/0.71	0.54/0.64	0.57/0.66	500/1500
Weighted avg	0.72/0.75	0.73/0.76	0.72/0.75	500/1500

Table 5-6 The classification report of MNB model (no “neutral” class)

N=200	Precision	Recall	F1-score	Support
Negative	0.93/0.95	0.98/0.97	0.95/0.96	335/959
Positive	0.65/0.71	0.38/0.54	0.48/0.61	40/113
Micro avg	0.91/0.93	0.91/0.93	0.91/0.93	375/1072
Macro avg	0.79/0.83	0.68/0.76	0.71/0.79	375/1072
Weighted avg	0.90/0.92	0.91/0.93	0.90/0.92	375/1072

For item 3:

For sentiment analysis, the majority class is “negative” which has 959 instances in the training set, and in the test set, the “negative” class has 335 instances. So the majority class baseline is (959/1072 + 335/375) / 2 = 0.894. With respect to VADER baseline, we get the accuracy is 0.480. Compared with standard models, the results are shown below (the accuracy of standard models):

Table 5-7 The results of standard models, Majority class and VADER (have "neutral" class)

DT	BNB	MNB	Majority class	VADER
0.688	0.716	0.738	0.655	0.430

Table 5-8 The results of standard models, Majority class and VADER (no "neutral" class)

DT	BNB	MNB	Majority class	VADER
0.901	0.893	0.904	0.894	0.480

For item 4:

Table 5-9 The metrics of DT model (top 200 words) of sentiment analysis (have "neutral" class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.688/0.699	0.470/0.575	0.420/0.466	0.416/0.479
With preprocessing	0.678/0.691	0.390/0.452	0.386/0.414	0.371/0.403

Table 5-10 The metrics of DT model (top 200 words) of sentiment analysis (no "neutral" class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.901/0.906	0.763/0.784	0.593/0.616	0.624/0.654
With preprocessing	0.893/0.895	0.447/0.447	0.50/0.50	0.472/0.472

Table 5-11 The metrics of BNB model of sentiment analysis (have "neutral" class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.716/0.832	0.471/0.590	0.410/0.558	0.400/0.561
With preprocessing	0.712/0.835	0.436/0.579	0.429/0.563	0.421/0.562

Table 5-12 The metrics of BNB model of sentiment analysis (no "neutral" class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.893/0.895	0.447/0.447	0.50/0.50	0.472/0.472
With preprocessing	0.893/0.895	0.447/0.447	0.50/0.50	0.472/0.472

Table 5-13 The metrics of MNB of sentiment analysis (have “neutral” class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.738/0.937	0.641/0.959	0.523/0.818	0.536/0.866
With preprocessing	0.744/0.925	0.591/0.945	0.546/0.809	0.552/0.857

Table 5-14 The metrics of MNB of sentiment analysis (no “neutral” class)

	Accuracy	Macro-p	Macro-r	Macro-f1
Without preprocessing	0.904/0.959	0.805/0.978	0.583/0.805	0.614/0.868
With preprocessing	0.899/0.962	0.732/0.973	0.646/0.822	0.675/0.880

Comparison:

From above tables, it is clear to see that removing neutral tweets from both training and test sets do increase the metrics value significantly for both positive and negative classes.

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 brief comparison of your methods in relation to the standard methods and the baselines.

Answer:

For both sentiment and topics analysis:

Based on the results of item 2 and item 3, I choose MNB algorithm as it has the highest metrics values and it costs the least time. And according to the results of item 4, adopting stop words removal and Porter stemming can improve the performance of the MNB.

For sentiment analysis:

Then we choose the optimal number of features:

Table 6-1 The accuracy of different N for sentiment analysis

N	Accuracy
100	0.702
200	0.728
300	0.728
400	0.720
All	0.744

From lots of experiments, I found that when we use all the features, then we could get the best result.

Next, I have tried some different stemming approaches (from nltk) that use all the features, and here is the result:

Table 6-2 The accuracy of different Stemmer for sentiment analysis

Stemmer	Accuracy
Porter Stemmer	0.744
Snowball Stemmer	0.742
Lancaster Stemmer	0.738

From the table, we can get a conclusion that Porter Stemmer is the best Stemmer for sentiment analysis in this assignment.

Finally, I doubt if lemmatization would be useful for this model, so I use it and compare it with the model without using lemmatization (both models use Porter Stemmer), here is the result:

Table 6-3 The accuracy of two models with and without lemmatization for sentiment analysis

	Accuracy
use lemmatization	0.742
No lemmatization	0.744

To sum up, for sentiment analysis, I choose to use MNB algorithm and use stop words removal and Porter Stemmer for the preprocessing, using all the features during the training.

For topic analysis:

I follow the same steps, and we can get another three tables:

Table 6-4 The accuracy of different N for topic analysis

N	Accuracy
100	0.386
200	0.432
300	0.448
400	0.426
500	0.450
600	0.452
700	0.434
800	0.444
900	0.434
1000	0.434
All	0.378

From the table, we decide to choose N = 600.

Table 6-5 The accuracy of different Stemmer for topic analysis

Stemmer	Accuracy
Porter Stemmer	0.452
Snowball Stemmer	0.436
Lancaster Stemmer	0.448

According to this table, we decide to use Porter Stemmer.

Table 6-6 The accuracy of two models with and without lemmatization for topic analysis

	Accuracy
use lemmatization	0.450
No lemmatization	0.452

Based on this table, we not choose lemmatization.

In conclusion, for topic analysis, I choose to use MNB algorithm and use stop words removal and Porter Stemmer for the preprocessing, using 600 features during the training.

And my own models' performance is better than that of the standard models, and their metrics values are above the baselines.