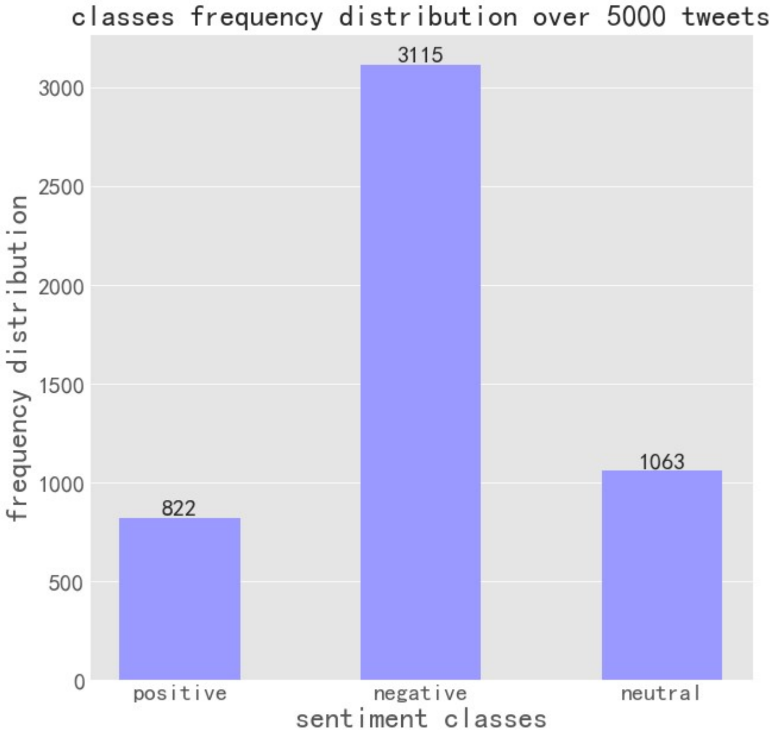
Assement2 Report

z5222381 Jizhou Liu

1.(1 mark) Give simple descriptive statistics showing the frequency distribution for the sentiment classes for the whole dataset of 5000 tweets. What do you notice about the distribution?

Figure 1　5000 tweets frequency



More than half of 5000 tweets are negative, and the number of neutral tweets is a little higher than the number of positive ones.

2. (2 marks) Develop BNB and MNB models from the training set using (a) the whole vocabulary, and (b) the most frequent 1000 words from the vocabulary (as defined using CountVectorizer, after preprocessing by removing "junk" characters). Show all metrics on

the test set comparing the two approaches for each method. Explain any similarities and differences in results.

Figure 2　BNB and MNB chose1000 words

			precision	recall	f1-score	accuracy
BNB	whole	macro avg	0.83	0.51	0.55	0.726
		weighted avg	0.77	0.73	0.68	
	1000	macro avg	0.71	0.74	0.72	0.784
		weighted avg	0.79	0.78	0.79	
MNB	whole	macro avg	0.80	0.59	0.64	0.762
		weighted avg	0.78	0.76	0.73	
	1000	macro avg	0.73	0.70	0.71	0.79
		weighted avg	0.78	0.79	0.78	

As we can see from the figure2, using the most frequent 1000 words from the vocabulary improve the accuracy. The accuracy of BNB change from 0.726 to 0.784, meanwhile the accuracy of MNB also improve by 0.03.

3. (2 marks) Evaluate the three standard models with respect to the VADER baseline. Show all metrics on the test set and comment on the performance of the baseline and of the models relative to the baseline.

Figure 3　VADER baseline

	precision	recall	f1-score	accuracy

VEDER	0.53	0.59	0.50	0.53
DT	0.63	0.54	0.57	0.70
BNB	0.79	0.44	0.45	0.688
MNB	0.78	0.54	0.58	0.737

The accuracy of VADER baseline is much lower than other three standard models. Because the test data does not have many emojis, and VADER performs better when dealing with emojis.

4. (2 marks) Evaluate the effect of preprocessing the input features by applying NLTK English stop word removal then NLTK Porter stemming on classifier performance for the three standard models. Show all metrics with and without preprocessing on the test set and explain the results.

Figure 4　NLTK

			precision	recall	f1-score	accuracy
DT	NLTK	macro avg	0.64	0.60	0.61	0.70
		weighted avg	0.69	0.70	0.69	
	standard	macro avg	0.63	0.54	0.57	0.70
		weighted avg	0.67	0.70	0.67	
BNB	NLTK	macro avg	0.77	0.48	0.51	0.70
		weighted avg	0.73	0.70	0.64	
	standard	macro avg	0.79	0.44	0.45	0.688

		weighted avg	0.73	0.69	0.61	
MNB	NLTK	macro avg	0.80	0.59	0.64	0.762
		weighted avg	0.76	0.76	0.73	
	standard	macro avg	0.77	0.60	0.64	0.76
		weighted avg	0.76	0.74	0.69	

After applying NLTK English stop word removal then NLTK Porter stemming, the test data set is better which can give good performance for the three standard models. The accuracy of three models get a little bit higher, and precision of three models obviously improved.

5. (2 marks) Evaluate the effect that converting all letters to lower case has on classifier performance for the three standard models. Show all metrics with and without conversion to lower case on the test set and explain the results.

Figure 5 lower case

			precision	recall	f1-score	accuracy
DT	lower case	macro avg	0.64	0.58	0.60	0.71
		weighted avg	0.68	0.71	0.69	
	standard	macro avg	0.63	0.54	0.57	0.70
		weighted avg	0.67	0.70	0.67	
BNB	lower	macro avg	0.83	0.51	0.55	0.726

	case	weighted avg	0.77	0.73	0.68	
	standard	macro avg	0.79	0.44	0.45	0.688
		weighted avg	0.73	0.69	0.61	
MNB	lower case	macro avg	0.80	0.59	0.64	0.762
		weighted avg	0.78	0.76	0.73	
	standard	macro avg	0.78	0.54	0.58	0.737
		weighted avg	0.76	0.74	0.69	

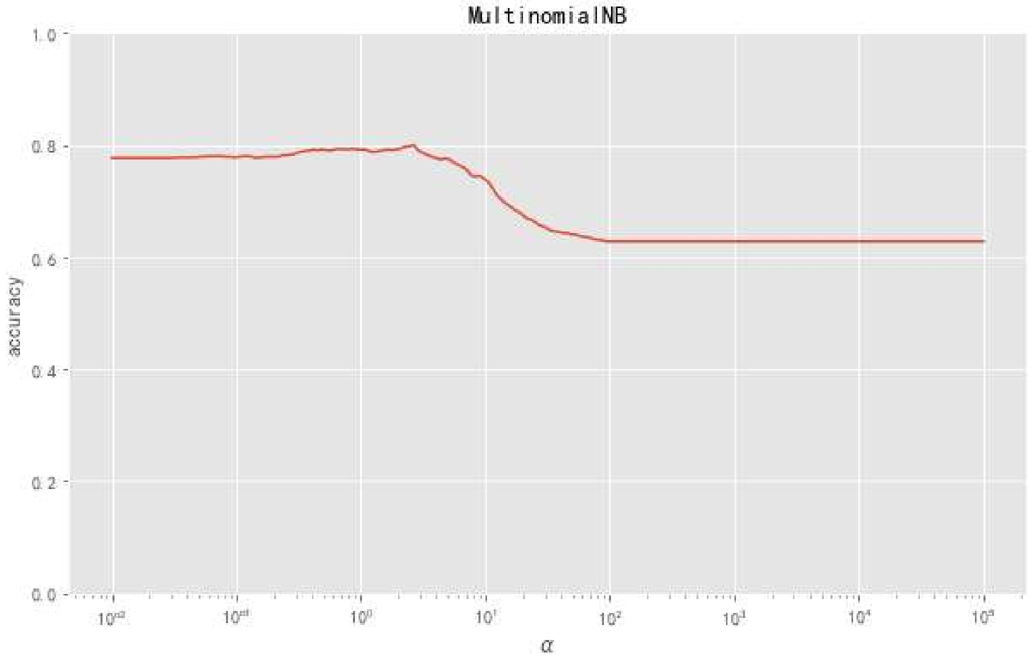
After converting all letters to lower case, the classifier performance for the three standard models gets better. The accuracy of three standard models improved by 0.01, 0.038, 0.025 separately.

6. (6 marks) Describe your best method for sentiment analysis and justify your decision. Give some experimental results for your method trained on the training set of 4000 tweets and tested on the test set of 1000 tweets. Provide a brief comparison of your model to the standard models and the baseline (use the results from the previous questions).

As we can see from the classification performance of three standard model, MNB model has the highest accuracy. Then apply NLTK English stop word removal then NLTK Porter stemming, convert all letters to lower case, the most frequent 1000 words from the vocabulary for test.

I also choose the alpha in MNB model to get the best performance. The accuracy of MNB changes as alpha changes, shown in the figure5 below:

Figure 6　accuracy with different alpha



I chose alpha = 2.67, at this point MNB model reach highest accuracy which is 0.8. The result can be seen in figure 7.

Figure 7　performance of my model

MNB	precision	recall	f1-score	accuracy
macro avg	0.76	0.71	0.73	0.80
weighted avg	0.79	0.80	0.79	