COMP9414 Assignment 2 Report

Yang Li Z3451444

Question 1

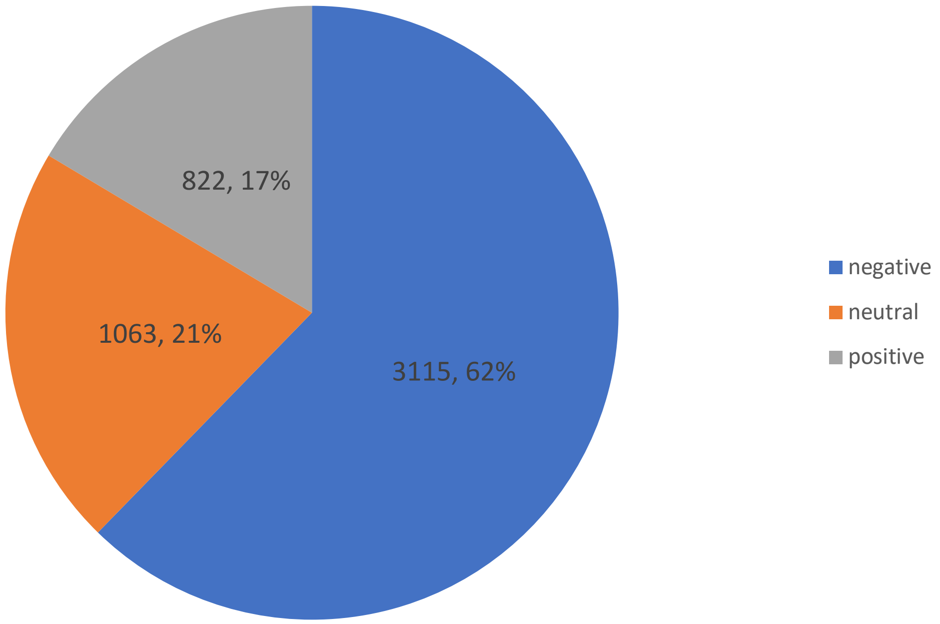

Figure 1 Sentiment distribution of 5000 tweets

There are three types of sentiments (negative, neutral and positive). Out of 5000 tweets, majority of people (62%) gives negative feedback.

Question 2

		BNB (all words)			
		precision	recall	f1-score	support
negative	all words	0.68	0.99	0.80	628
	1000 words	0.87	0.83	0.85	628
neutral	all words	0.77	0.21	0.33	210
	1000 words	0.60	0.67	0.63	210
positive	all words	0.91	0.12	0.22	162
	1000 words	0.61	0.65	0.63	162

Table 1 BNB metrics comparison between using all words and most frequent 1000 words categorized by sentiments

		MNB (all words)			
		precision	recall	f1-score	support
negative	all words	0.72	0.99	0.84	628
	1000 words	0.84	0.89	0.86	628
neutral	all words	0.79	0.26	0.39	210
	1000 words	0.63	0.54	0.58	210

positive	all words	0.83	0.39	0.53	162
	1000 words	0.67	0.62	0.65	162

Table 2 MNB metrics comparison between using all words and most frequent 1000 words categorized by sentiments

The tables above show the metrics comparison between using all words and only the most frequently used 1000 words categorized by sentiments for BNB and MNB models. It is clear that the negative sentiment has higher precision, recall and f1-score than the other sentiments for all models. This is due to fact that the negative sentiment has the largest distribution over the training set, so that it is able to come up with better predictions.

	BNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
all words	0.69	0.79	0.44	0.45
1000 words	0.76	0.69	0.71	0.70

Table 3 BNB metrics

	MNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
all words	0.74	0.78	0.54	0.58
1000 words	0.77	0.71	0.69	0.70

Table 4 MNB metrics

Configuration notice
Accuracy is the same as micro-precision, micro-recall, micro-f1 score and accuracy.

Comparisons
<u>Micro precision vs macro precision</u>
For both models, macro-precision is higher than micro-precision. As micro-precision is more focused on the majority class (negative sentiment in this case), and from Table 1 we know that negative sentiment has better prediction result. Therefore, micro-precision has higher value.

<u>Using all words vs using the most frequent 1000 words</u>
Using the most frequent 1000 words has better accuracy, macro-recall and macro-f1 score than using all words.

<u>BNB vs MNB</u>
For the current testing set, MNB has a small advantage over BNB.

Question 3

	VADAR			
	precision	recall	f1-score	support
negative	0.91	0.48	0.63	628
neutral	0.36	0.43	0.39	210
positive	0.34	0.89	0.49	162

Table 5 metrics categorized by sentiments (VADAR)

Compared above table with Table 1, it is noticeable that VADAR is not performing as good as BNB and MNB, especially for the negative sentiments which is the majority class of the testing data.

	accuracy	macro-precision	macro-recall	macro-f1 score
VADAR	0.54	0.54	0.60	0.51
DT	0.70	0.62	0.54	0.56
BNB	0.69	0.79	0.44	0.45
MNB	0.74	0.78	0.54	0.58

Table 6 metrics comparison between VADAR, DT, BNB and MNB (use all words)

Configuration notice
accuracy is the same as micro-precision, micro-recall, micro-f1 score and accuracy.

Comparisons
The table above shows the metrics comparison between VADAR, DT, BNB and MNB. Generally speaking, our models (DT, BNB and MNB) has better performance than VADAR, and have a higher accuracy, macro-precision and macro-f1 score. Here are the possible causes:

- VADAR is a crowd sourcing software that could be highly unreliable
- VADAR is a trained model without using our training data set, hence it is not fine-tuned to these tweet texts.
- VADAR performs really well with emojis and slangs, whereas our tweet texts are mostly normal sentences.

Question 4

Configuration notice
accuracy is the same as micro-precision, micro-recall, micro-f1 score and accuracy.
Processed means removing stop words and stemming.

DT Metrics

		DT			
		precision	recall	f1-score	support
negative	normal	0.73	0.90	0.81	628
	processed	0.77	0.85	0.81	628
neutral	normal	0.46	0.25	0.33	210
	processed	0.43	0.38	0.41	210
positive	normal	0.68	0.48	0.56	162
	processed	0.70	0.56	0.62	162

Table 7 DT metrics comparison with and without pre-processing categorized by sentiments

	DT			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.70	0.62	0.54	0.56
processed	0.70	0.64	0.59	0.61

Table 8 DT metrics comparison (with and without pre-processing)

Comments on the metrics:
DT model metrics improved slightly after removing stop words and stemming. More specifically, accuracy remained the same, small increase of macro-precision, macro-recall and macro-f1 score.

BNB Metrics

	BNB				
		precision	recall	f1-score	support
negative	normal	0.68	0.99	0.80	628
	processed	0.69	0.98	0.81	628
neutral	normal	0.77	0.21	0.33	210
	processed	0.74	0.23	0.35	210
positive	normal	0.91	0.12	0.22	162
	processed	0.88	0.23	0.36	162

Table 9 BNB metrics comparison with and without pre-processing categorized by sentiments

	BNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.69	0.79	0.44	0.45
processed	0.70	0.77	0.48	0.51

Table 10 BNB metrics comparison (with and without pre-processing)

Comments on the metrics:
BNB model metrics has no significant improvement after removing stop words and stemming. More specially, accuracy and macro-precision decreased for a small amount, but there are some increments in macro-recall and macro-f1 score.

MNB Metrics

	MNB				
		precision	recall	f1-score	support
negative	normal	0.72	0.99	0.84	628
	processed	0.75	0.97	0.85	628
neutral	normal	0.79	0.26	0.39	210
	processed	0.76	0.32	0.45	210
positive	normal	0.83	0.39	0.53	162
	processed	0.78	0.49	0.60	162

Table 11 MNB metrics comparison with and without pre-processing categorized by sentiments

	MNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.74	0.78	0.54	0.58
processed	0.76	0.77	0.59	0.64

Table 12 MNB metrics comparison (with and without pre-processing)

Comments on the metrics:
MNB model metrics has some improvements after removing stop words and stemming. More specifically, accuracy has increased, so as the macro-recall and macro-f1 score. However, there is a small decrease in macro-precision.

Comparisons
Generally speaking, performance has improved slightly for each model after removing the stop words and stemming. Although it is noticeable that some metrics stayed the same or even decreased after applying stop words removal and stemming.

Possible causes
Tweets are short and simple sentences; thus, key words that represent sentiments in the sentence are usually the same part of speech (e.g. verb, adjective, adverb). This might explain why there is no significant improvement even after removing stop words and stemming. In addition, there is no case conversion at the moment, so same word with different casing would be considered as two different words.

Question 5

DT Metrics

	DT				
		precision	recall	f1-score	support
negative	normal	0.73	0.90	0.81	628
	lower case	0.75	0.89	0.81	628
neutral	normal	0.46	0.25	0.33	210
	lower case	0.49	0.28	0.35	210
positive	normal	0.68	0.48	0.56	162
	lower case	0.68	0.71	0.68	162

Table 13 DT metrics comparison with and without lower case conversion categorized by sentiments

	DT			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.70	0.62	0.54	0.56
lower case	0.71	0.64	0.58	0.59

Table 14 DT metrics comparison (with and without lower case conversion)

<u>Comments on the metrics:</u>
DT model metrics improved slightly after converting all words to lower case. All metrics have small increment.

BNB Metrics

	BNB				
		precision	recall	f1-score	support
negative	normal	0.68	0.99	0.80	628
	lower case	0.71	0.99	0.83	628
neutral	normal	0.77	0.21	0.33	210
	lower case	0.83	0.32	0.46	210
positive	normal	0.91	0.12	0.22	162
	lower case	0.96	0.27	0.42	162

Table 15 BNB metrics comparison with and without lower case conversion categorized by sentiments

	BNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.69	0.79	0.44	0.45
lower case	0.73	0.83	0.52	0.57

Table 16 BNB metrics comparison (with and without lower case conversion)

Comments on metrics:
BNB model metrics has noticeable improvement after converting all words to lower case. All metrics have some increases.

MNB Metrics

		MNB			
		precision	recall	f1-score	support
negative	normal	0.72	0.99	0.84	628
	lower case	0.75	0.98	0.85	628
neutral	normal	0.79	0.26	0.39	210
	lower case	0.81	0.33	0.47	210
positive	normal	0.83	0.39	0.53	162
	lower case	0.83	0.46	0.60	162

Table 17 MNB metrics comparison with and without lower case conversion categorized by sentiments

	MNB			
	accuracy	macro-precision	macro-recall	macro-f1 score
normal	0.74	0.78	0.54	0.58
processed	0.76	0.80	0.59	0.64

Table 18 MNB metrics comparison (with and without lower case conversion)

Configuration notice
Case conversion is applied to both the training set and the testing set.

Comments on the metrics:
MNB model metrics has noticeable improvement after converting all words to lower case. All metrics have some increases.

Comparisons
The case conversion has very positive result, performance has improved for all metrics in each model.

Possible causes
Case conversion has made the training and testing tweet texts for uniformed. For example, in the original data set, “Happy” and “happy” are actually in two groups. But after the case conversion, they are now merged as one. This would definitely improve the accuracy of the model as words are categorized into larger clusters.

Question 6

Parameters chosen for the best method

Model chosen	MNB
Max features	2000 words
Lower case conversion	Enabled
Remove stop words	Disabled
Stemming	Disabled

Reason of choices

<u>Model:</u> MNB model has better performance than BNB and DT.
<u>Max features:</u> I have tested max features from 500 words to 3000 words, and the test result shows that 2000 words would maximize the performance.
<u>Lower case conversion:</u> This feature is enabled as it is proved in the previous question that it will improve all metrics for all models.
<u>Remove stop words and stemming:</u> I have tested only removing stop words, only stemming, or applying both. However, results have shown that the metrics actually dropped for a small amount comparing to not applying these two features.

Conclusion

Pre-processing plays an important role in the supervised training process. The quality of the data largely determines how effective the model is. On top of that, the more similarities between the training data and testing data, the better the performance is.

Performance comparison

		precision	recall	f1-score	support
negative	VADAR	0.91	0.48	0.63	628
	DT	0.73	0.90	0.81	628
	BNB	0.68	0.99	0.80	628
	MNB	0.72	0.99	0.84	628
	Best method	**0.85**	**0.93**	**0.89**	**628**
neutral	VADAR	0.36	0.43	0.39	210
	DT	0.46	0.25	0.33	210
	BNB	0.77	0.21	0.33	210
	MNB	0.79	0.26	0.39	210
	Best method	**0.71**	**0.55**	**0.62**	**210**
positive	VADAR	0.34	0.89	0.49	162
	DT	0.68	0.48	0.56	162
	BNB	0.91	0.12	0.22	162
	MNB	0.83	0.39	0.53	162
	Best method	**0.75**	**0.69**	**0.72**	**162**

Table 19 Metrics comparison between models categorized by sentiments

	accuracy	macro-precision	macro-recall	macro-f1 score
VADAR	0.54	0.54	0.60	0.51
DT	0.70	0.62	0.54	0.56

BNB	0.69	0.79	0.44	0.45
MNB	0.74	0.78	0.54	0.58
Best method	**0.81**	**0.77**	**0.72**	**0.74**

Table 20 metrics comparison between models

Comparisons

As shown in the above two tables, our best method outperforms the baseline (VADAR), DT, BNB, MNB in most metrics.