KE5204 NEW MEDIA AND SENTIMENT MINING

PROGRAMMING ASSIGNMENT

A summary of sentiment classification modelling for restaurant reviews from Yelp

SUBMISSION DATE: 28 OCTOBER 2018

GROUP 04
SIDDHARTH PANDEY
PRANSHU RANJAN SINGH
NYON YAN ZHENG
TAN KOK KENG

INSTITUTE OF SYSTEMS SCIENCE
NATIONAL UNIVERSITY OF SINGAPORE

Data Collection

We have scraped 2,472 restaurant reviews from the Yelp website. The scraped reviews consist of 46 unique restaurants from 18 different cuisines.

Data Description

We picked 4 restaurants each from 4 different cuisines as the test data and the remaining reviews are used as training data. Yelp reviews with ratings 1 to 3 are tagged as "negative" and ratings 4 and 5 are tagged as "positive".

Training Data

The training data from Yelp is combined with the standard reviews set from the workshop. Table 1 shoes the breakdown of training data by sentiment.

Table 1 Breakdown of training data by sentiment

Cuisine	Positive	Negative	Total Reviews Count
From Yelp	1,336 (63%)	772 (37%)	2,108
Standard set	9,318 (46%)	11,011 (54%)	20,329
Total	**10,654 (47%)**	**11,783 (53%)**	**22,437**

Test Data

Table 2 shoes the breakdown of the test dataset by cuisine and Table 3 shows the breakdown by sentiment.

Table 2 Breakdown of test data by cuisine

Cuisine	Restaurant Count	Reviews Count
Asian Fusion	4	63
French	4	137
Thai	4	71
Vietnamese	4	93
Total	**16**	**364**

Table 3 Breakdown of test data by sentiment

Positive	Negative	Total Reviews Count
238 (65%)	126 (37%)	364

Data Pre-processing

We have done the following pre-preprocessing:

1. Tokenize words
2. Lemmatize words according to the POS tag

In one of the model, we have also used the negation feature where the words after a negative text are tagged with "NEG".

We did not use case lowering to preserve the sentiment in capitalised words.

Most Informative Words

Using the Naïve Bayes Classifier from NLTK, the most informative words are obtained and shown below.

```
Most Informative Features
                  beggar = True                1 : -1     =     66.0 : 1.0
                 chooser = True                1 : -1     =     65.3 : 1.0
                   rib- = True                 1 : -1     =     64.5 : 1.0
                  water- = True                1 : -1     =     64.5 : 1.0
               preoccupy = True               -1 : 1      =     54.0 : 1.0
               drinkable = True               -1 : 1      =     54.0 : 1.0
                 orchard = True               -1 : 1      =     53.3 : 1.0
                Terrible = True               -1 : 1      =     53.3 : 1.0
                 Orchard = True                1 : -1     =     52.0 : 1.0
                     Pho = True               -1 : 1      =     48.4 : 1.0
```

Summary of Modelling Results

We have tried Naïve Bayes (NB), Random Forest(RF) and Support Vector Classifier (SVC) models in this assignment. We have also attempted using both unigrams and bigrams as the features, however, this did not improve our precision score. We have also tried the SVC model with the negation features and this improves our precision by 1 percentage point (from 0.83 to 0.84). Furthermore, we tried the SVC using a linear kernel and the test results did not improve. This could be due to overfitting of training data where the accuracy score is 1.00.

The summary results are shown in the table below.

Table 4 Summary of Training and Test Scores

Model	Train			Test		
	Accuracy	Weighted Precision	Weighted Recall	Accuracy	Weighted Precision	Weighted Recall
NLTK Naïve Bayes Classifier (unigrams)	0.83	0.86	0.83	0.65	0.78	0.65
Random Forest (unigrams)	0.86	0.86	0.86	0.72	0.77	0.72
SVC (unigrams, rbf kernel)	0.93	0.93	0.93	0.82	0.83	0.82
SVC (unigrams + bigrams, rbf kernel)	0.86	0.87	0.86	0.74	0.79	0.74
SVC (unigrams, rbf kernel, negation features)	**0.94**	**0.94**	**0.94**	**0.83**	**0.84**	**0.83**
SVC (unigrams, linear kernel, negation features)	1.00	1.00	1.00	0.73	0.76	0.73

In summary, the RF model performed better than the NB model in terms of accuracy but still underperform all the SVC models. Amongst the SVC models, we found that the model using the rbf kernel with negation features gave the overall best accuracy score of 0.83 on the test set as highlighted in the table above. This model also gives the highest weighted precision of 0.84 for both the positive and negative labels.

The classification reports and confusion matrices for all the 6 models are provided in the appendices.

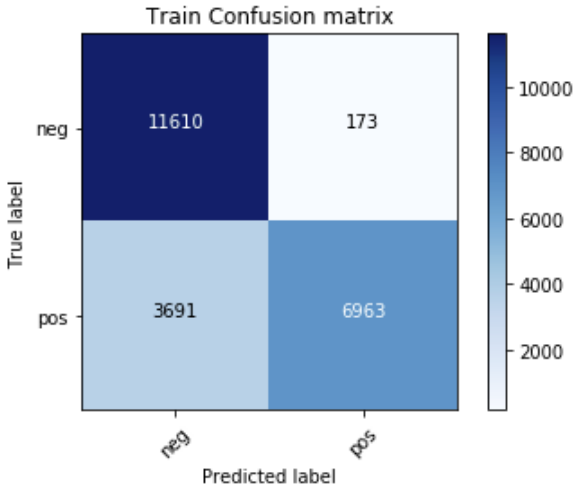
Appendix A – Classification Report for Naïve Bayes Classifier

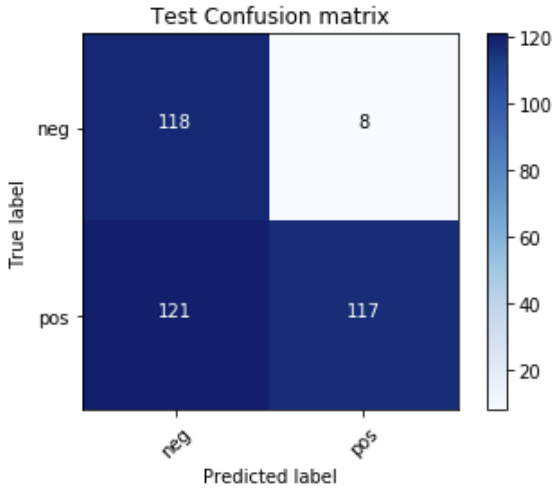
Train Results

	precision	recall	f1-score	support
Neg	0.76	0.99	0.86	11783
Pos	0.98	0.65	0.78	10654
micro avg	0.83	0.83	0.83	22437
macro avg	0.87	0.82	0.82	22437
weighted avg	0.86	0.83	0.82	22437

Test Results

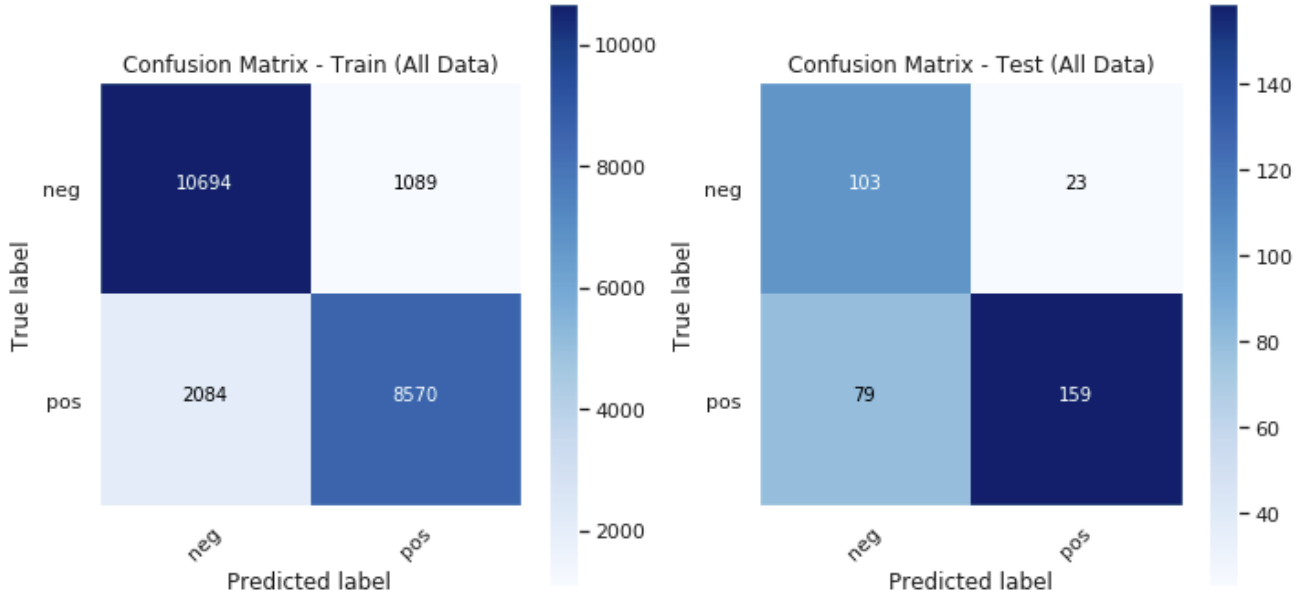
	precision	recall	f1-score	support
Neg	0.49	0.94	0.65	126
Pos	0.94	0.49	0.64	238
micro avg	0.65	0.65	0.65	364
macro avg	0.71	0.71	0.65	364
weighted avg	0.78	0.65	0.65	364





Appendix B – Classification Report for Random Forest Classifier

```
train accuracy:  0.8585818068369212
test accuracy:  0.7197802197802198
train report:                precision    recall  f1-score   support

           0       0.84      0.91      0.87     11783
           1       0.89      0.80      0.84     10654

   micro avg       0.86      0.86      0.86     22437
   macro avg       0.86      0.86      0.86     22437
weighted avg       0.86      0.86      0.86     22437

test report:                precision    recall  f1-score   support

           0       0.57      0.82      0.67       126
           1       0.87      0.67      0.76       238

   micro avg       0.72      0.72      0.72       364
   macro avg       0.72      0.74      0.71       364
weighted avg       0.77      0.72      0.73       364
```



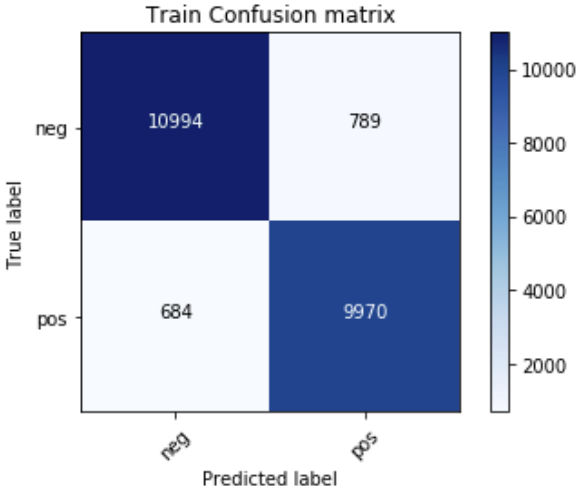
Appendix C – Classification Report for Support Vector Classifier (unigrams, rbf kernel)

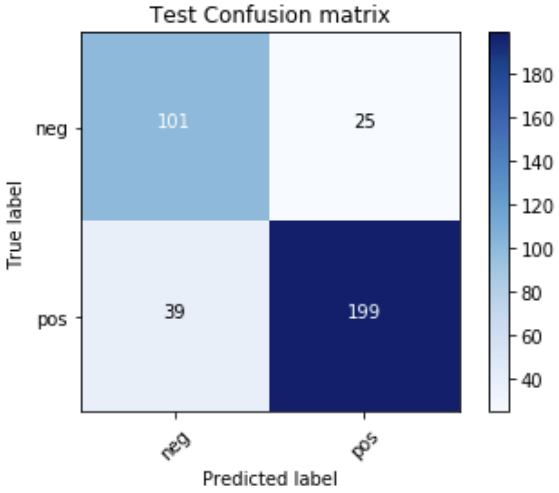
Train Results

	precision	recall	f1-score	support
Neg	0.94	0.93	0.94	11783
Pos	0.93	0.94	0.93	10654
micro avg	0.93	0.93	0.93	22437
macro avg	0.93	0.93	0.93	22437
weighted avg	0.93	0.93	0.93	22437

Test Results

	precision	recall	f1-score	support
Neg	0.72	0.80	0.76	126
Pos	0.89	0.84	0.86	238
micro avg	0.82	0.82	0.82	364
macro avg	0.80	0.82	0.81	364
weighted avg	0.83	0.82	0.83	364





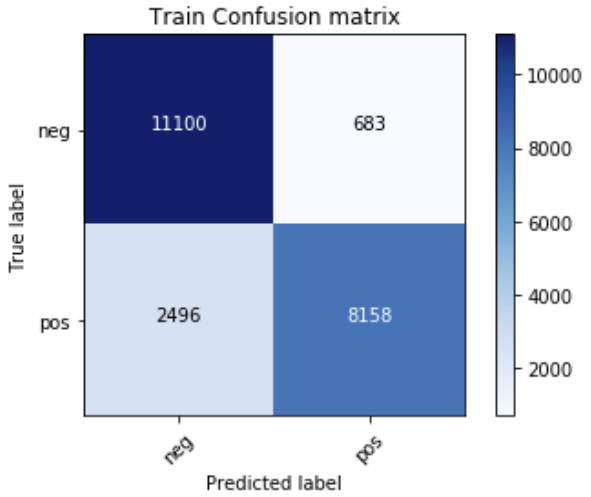
Appendix D – Classification Report for Support Vector Classifier (unigrams + bigrams, rbf kernel)

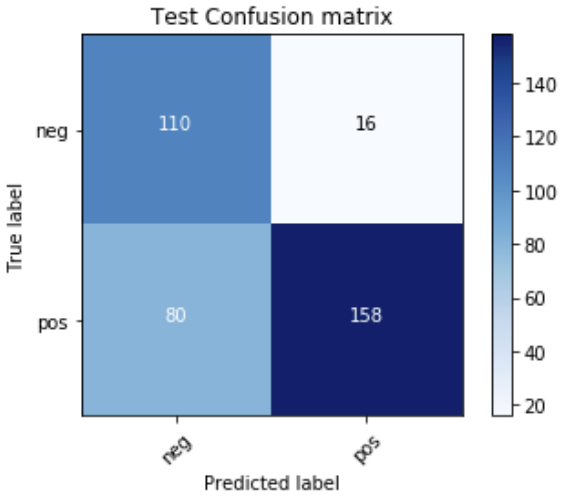
Train Results

	precision	recall	f1-score	support
Neg	0.82	0.94	0.87	11783
Pos	0.92	0.77	0.84	10654
micro avg	0.86	0.86	0.86	22437
macro avg	0.87	0.85	0.86	22437
weighted avg	0.87	0.86	0.86	22437

Test Results

	precision	recall	f1-score	support
Neg	0.58	0.87	0.70	126
Pos	0.91	0.66	0.77	238
micro avg	0.74	0.74	0.74	364
macro avg	0.74	0.77	0.73	364
weighted avg	0.79	0.74	0.74	364





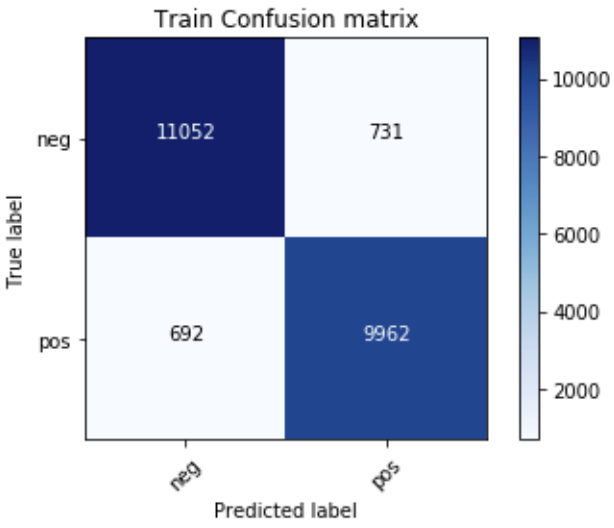
Appendix E – Classification Report for Support Vector Classifier (unigrams, rbf kernel, negation features)

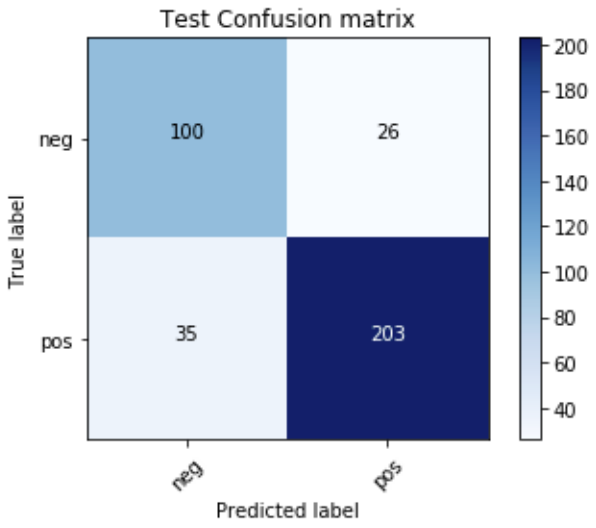
Train Results

	precision	recall	f1-score	support
Neg	0.94	0.94	0.94	11783
Pos	0.93	0.94	0.93	10654
micro avg	0.94	0.94	0.94	22437
macro avg	0.94	0.94	0.94	22437
weighted avg	0.94	0.94	0.94	22437

Test Results

	precision	recall	f1-score	support
Neg	0.74	0.79	0.77	126
Pos	0.89	0.85	0.87	238
micro avg	0.83	0.83	0.83	364
macro avg	0.81	0.82	0.82	364
weighted avg	0.84	0.83	0.83	364





Appendix F – Classification Report for Support Vector Classifier (unigrams, linear kernel, negation features)

Train Results

	precision	recall	f1-score	support
Neg	1.00	1.00	1.00	11783
Pos	1.00	1.00	1.00	10654
micro avg	1.00	1.00	1.00	22437
macro avg	1.00	1.00	1.00	22437
weighted avg	1.00	1.00	1.00	22437

Test Results

	precision	recall	f1-score	support
Neg	0.59	0.77	0.67	126
Pos	0.85	0.71	0.78	238
micro avg	0.73	0.73	0.73	364
macro avg	0.72	0.74	0.72	364
weighted avg	0.76	0.73	0.74	364