CS919 Exercise 2: Report

1532385

December 10, 2018

1 Introduction

Three different classifiers were built: logistic regression (LR), Naive Bayes (NB) and support vector machine (SVM).

2 Methodology

The tokeniser used was from the provided file twokenize.py. This was used because it performed slightly better than the build in tokenisers and better than the NLTK tokeniser.

For text preprocessing: emails, URLs and user mentions were replaced with separate identifiers. Also all standalone numbers were removed. This combination resulted in the highest accuracy. Removing non-alphanumeric characters resulted in a decrease in accuracy, hence they were kept.

For features: count vectorizer function was used and the length of the tweet.

3 Results

The results show the accuracy achieved from training data, then the accuracy of each test data with the corresponding F1 score. Also a confusion matrix is shown. After each confusion matrix is shown the average model accuracy is displayed.

3.1 Classifier 1: Logistic Regression

Training accuracy: 0.6917960088691796

Overall accuracy 1: 0.6527895780232229
F1 score = 0.572

Table 1: myclassifier1 text 1

	positive	negative	neutral
positive	0.714	0.050	0.236
negative	0.134	0.732	0.134
neutral	0.243	0.158	0.599

Overall accuracy 2: 0.6788990825688074
F1 score = 0.591

Table 2: myclassifier1 text 2

	positive	negative	neutral
positive	0.783	0.051	0.166
negative	0.173	0.654	0.173
neutral	0.320	0.102	0.578

Overall accuracy 3: 0.6305170239596469
F1 score: 0.555

Table 3: myclassifier1 text 3

	positive	negative	neutral
positive	0.754	0.049	0.196
negative	0.236	0.574	0.190
neutral	0.297	0.130	0.574

Average Accuracy: 0.6540685615172257

3.2 Classifier 2: NB

Training accuracy: 0.6518847006651884

Overall accuracy 1: 0.6046445766071934
F1 score = 0.494

Table 4: myclassifier2 text 1

	positive	negative	neutral
positive	0.606	0.093	0.301
negative	0.097	0.755	0.148
neutral	0.241	0.170	0.589

Overall accuracy 2: 0.6470588235294118
F1 score = 0.474

Table 5: myclassifier2 text 2

	positive	negative	neutral
positive	0.686	0.084	0.230
negative	0.095	0.619	0.286
neutral	0.296	0.119	0.584

Overall accuracy 3: 0.6128625472887768
F1 score = 0.501

Table 6: myclassifier2 text 3

	positive	negative	neutral
positive	0.662	0.086	0.252
negative	0.216	0.582	0.203
neutral	0.270	0.155	0.575

Average Accuracy: 0.6215219824751274

3.3 Classifier 3: SVM

Training accuracy: 0.6718403547671841

Overall accuracy 1: 0.6451430189747946
F1 score = 0.489

Table 7: myclassifier3 text 1

	positive	negative	neutral
positive	0.764	0.056	0.179
negative	0.132	0.806	0.062
neutral	0.253	0.173	0.574

Overall accuracy 2: 0.6616297895304911
F1 score = 0.497

Table 8: myclassifier3 text 2

	positive	negative	neutral
positive	0.808	0.062	0.131
negative	0.049	0.829	0.122
neutral	0.339	0.116	0.545

Overall accuracy 3: 0.6204287515762925
F1 score = 0.460

Table 9: myclassifier3 text 3

	positive	negative	neutral
positive	0.804	0.071	0.125
negative	0.170	0.755	0.074
neutral	0.308	0.149	0.543

Average Accuracy: 0.6424005200271928

3.4 Summary

Table 10: Summary of results

	LR	NB	SVM
Average Accuracy	0.65	0.62	0.64

4 Discussion

From the results you can see that LR performed better than the other classifiers in every test set.

Overall, the final accuracy achieved could be much better.

This work can be improved further by extracting more features and using them to carry out a gridsearch. Also other classifiers should be tested such as decision trees and neural networks.