4. 1

d). The dimensionality of the feature matrix is:

dimensionality of the all the tweet: (630, 1140930)

dimensionality of training data: (560, 1014160)

dimensionality of test data: (70, 126770)

4.2

b). It is beneficial to maintain class proportions across folds since the amount of data for each model needs to be reasonable, in order to make the model learn a proper decision boundary. If not doing this, in some extreme case, the model may even not be aware of one of the class for classification.

C	**accuracy**	**F1-score**	**AUROC**
$\mathbf{10^{-3}}$	0.7089	0.8297	0.5
$\mathbf{10^{-2}}$	0.7107	0.8306	0.5031
$\mathbf{10^{-1}}$	0.8060	0.8755	0.7188
$\mathbf{10^{0}}$	0.8146	0.8749	0.7531
$\mathbf{10^{1}}$	0.8182	0.8766	0.7592
$\mathbf{10^{2}}$	0.8182	0.8766	0.7592
Best C	100	100	100

Parameter C is used to decide the boundary, and it could adjust larger slack variables. Larger value C would give a relevantly small margin to avoid misclassification. Smaller value of C would instead give a large margin which may include more misclassification. In our case, the larger C leads to much better performance for our model.

4.3

a). The best C value is 100.

c).

Performance Metric	**Linear-Kernel SVM score**
Accuracy	0.7429
F1_score	0.4375
AUROC	0.6259