

CS 4372

ASSIGNMENT 4

Names of students in your group:

James Hooper

Hritik Panchasara

Number of free late days used: 0

Note: You are allowed a total of 4 free late days for the entire semester. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

For model creation and pre-processing

<https://scikit-learn.org/>

For data manipulation

<https://numpy.org/>

<https://pandas.pydata.org/>

REPORT

- The results showcase that the trials with the highest results have the alpha set to 1, Fit Prior parameter set to False and having a Test Split of .1 (90% training data, 10% test data). When fit prior is set to false the better the accuracy. In the logs the first 25 predicted and y-test actual values are printed to be compared to one another. One interesting thing of note is that it seems as if the model mainly guesses the value of '0'. This could probably be fixed if the dataset was a bit more diverse in the sentiment value. In fact, this is represented in the final print of the different airline sentiments. The *airline with the highest average sentiment value* was Virgin America with about a .942 average sentiment value. This is in relation to the sentiment values: negative = 0, neutral = 1, positive = 2. So even the most highly rated airline (sentiment wise) does not even reach a neutral sentiment. It is shown that all airlines are perceived rather negatively. With this brief data showcase we can see why the model may be more likely to predict '0' for any given review. If we look at the printed confusion matrices for each trial, we can see that the value of '0' was predicted incorrectly way more than the other types of incorrect assumptions.
- The overall results from each trial however were not great. At worst we roughly got 68% accuracy and at best we roughly got 77% accuracy. Again, as mentioned previously if the dataset was more diverse (having more neutral and positive reviews) we have been able to achieve more results. Other than that, there may be other ways to tune the model to lead to better results and of course having more than 5 trials could lead to better conclusions.

LOGS

Trial 1

Test Size: 0.1

Alpha: 1

Fit Prior: True

X_train shape: (13176, 15051)

X_test shape: (1464, 15051)

Model Prediction Result:

0.6878415300546448

Confusion Matrix:

[[911 5 0]

[258 60 5]

[179 10 36]]

Predicted

[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 1 0]

Y_test

[1 2 0 1 1 1 2 0 0 0 0 0 1 0 0 2 0 2 1 1 0 0 0 0 1]

Trial 2

Test Size: 0.1

Alpha: 1

Fit Prior: False

X_train shape: (13176, 15051)

X_test shape: (1464, 15051)

Model Prediction Result:

0.7711748633879781

Confusion Matrix:

[[901 27 12]

[168 114 20]

[89 19 114]]

Predicted

[2 2 0 0 0 0 0 0 1 2 0 1 0 2 0 0 0 0 0 0 0 1 0 2 0]

Y_test

[2 1 0 1 0 0 0 0 0 2 1 1 0 2 0 0 2 0 0 0 2 0 0 2 2]

Trial 3

Test Size: 0.1

Alpha: 0.5

Fit Prior: True

X_train shape: (13176, 15051)

X_test shape: (1464, 15051)

Model Prediction Result:

0.7117486338797814

Confusion Matrix:

[[885 8 2]

[233 78 8]

[155 16 79]]

Predicted

[0 0 0 2 0 2 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0]

Y_test

[2 0 0 2 0 2 0 0 0 1 1 0 2 1 0 0 1 0 1 0 0 1 0 0 1]

Trial 4

Test Size: 0.1

Alpha: 0.5

Fit Prior: False

X_train shape: (13176, 15051)

X_test shape: (1464, 15051)

Model Prediction Result:

0.7657103825136612

Confusion Matrix:

[[826 50 22]

[150 151 35]

[59 27 144]]

Predicted

[0 0 1 0 0 0 0 2 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0]

Y_test

[1 0 1 0 0 2 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 2 0]

Trial 5

Test Size: 0.1
Alpha: 0.75
Fit Prior: True
X_train shape: (13176, 15051)
X_test shape: (1464, 15051)

Model Prediction Result:
0.7049180327868853

Confusion Matrix:
[[923 5 0]
[242 64 6]
[169 10 45]]

Predicted
[0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

Y_test
[0 0 0 0 1 0 0 1 1 1 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0]

Average Sentiment of Each Airline
Negative = 0
Neutral = 1
Positive = 2

	airline	avg_sentiment
0	virgin america	0.942460
1	united	0.439822
2	southwest	0.745455
3	delta	0.815032
4	us airways	0.315482
5	american	0.411381

Process finished with exit code 0