Analysis US Airline Sentiments and Classification Models Evaluations

Meisi Li (meli7912)
Jun 1, 2021

Research Problem

Dataset: Tweet dataset from Kaggle[1] (14640 rows x 15 columns)

Problem: Use word embedding to find the relationship of sentiment and feedback from Twitter

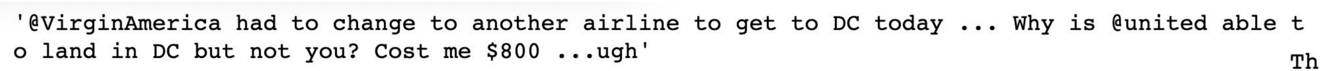
Hypothesis

- H0: cannot improve the performance with different model

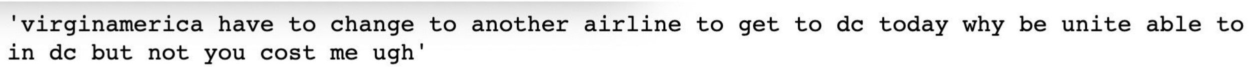
Dataset: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

Approach

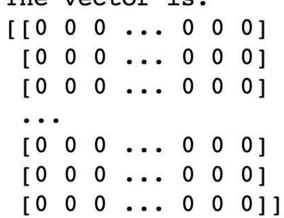
- Data Processing: clean data and CountVectorizer



```
'@VirginAmerica had to change to another airline to get to DC today ... Why is @united able t
o land in DC but not you? Cost me $800 ...ugh'
```



```
'virginamerica have to change to another airline to get to dc today why be unite able to
in dc but not you cost me ugh'
```



```
The vector is:
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

- Split data into training set and testing set: 0.33
- Testing set and validation set: 0.5
- Model selection - Grid Search
  - Baseline: naive Bayes
  - Decision Tree & Random Forest & logistic regression
  - Ensemble model

Evaluation

- f1_score, accuracy
  - Five models → best model

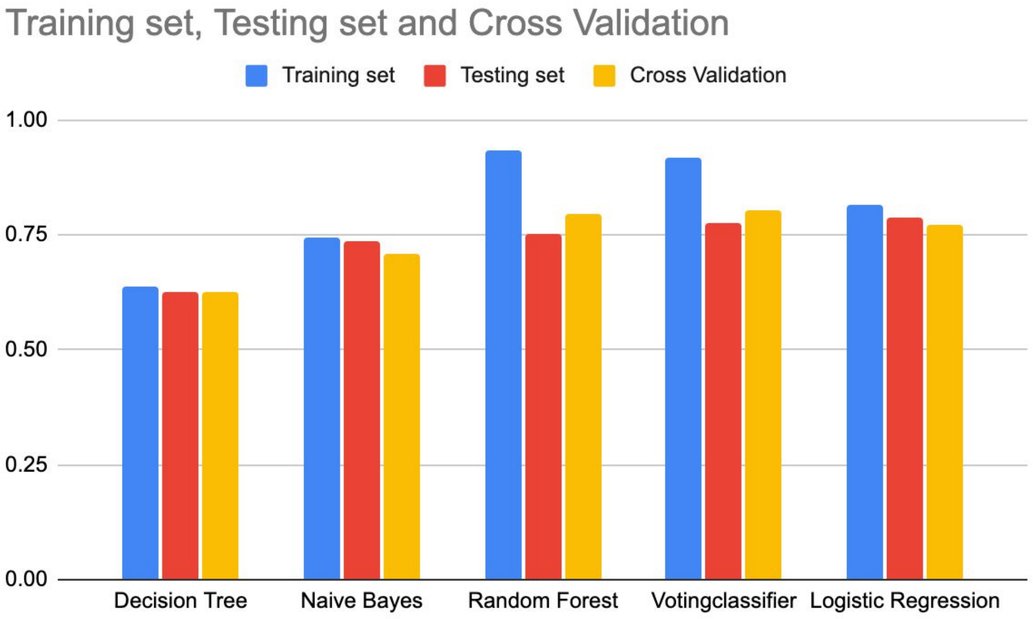
- McNemar’s test
  - Baseline compares with best model

Result

f1_score & accuracy

Best model:

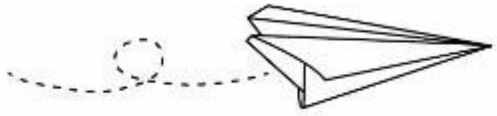
Voting Classifier



McNemar’s Test Result

P-value: 1.0670431364752282e-10 < 0.05

Conclusion: reject H0



Thank You