Assignment 7

Result

test 1

	without POS	POS
classification tree	**0.72, 0.7149**	0.6973, 0.6867

test 2

Compare the classification tree method with different limit of max features without POS tagging.

max_feature	without POS
500000	**0.72, 0.7149**
5000	0.716, 0.7232
3000	0.702, 0.6972
1000	0.706, 0.7036

test 3

model	without POS, max_feature = 500000
classification tree	0.72, 0.7149
random forest	0.839, 0.8338
neural network (default setting)	**0.842, 0.8381**
neural network (hidden_layer_sizes=(500, 2))	0.838, 0.8350

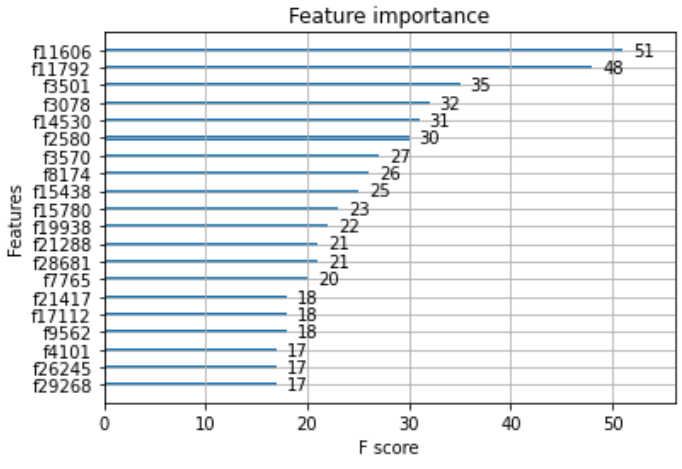
(optional) other classification algorithms

model	without POS, max_feature = MAX
XGBoost	0.803, 0.7979
DNN	**0.865, 0.8637**

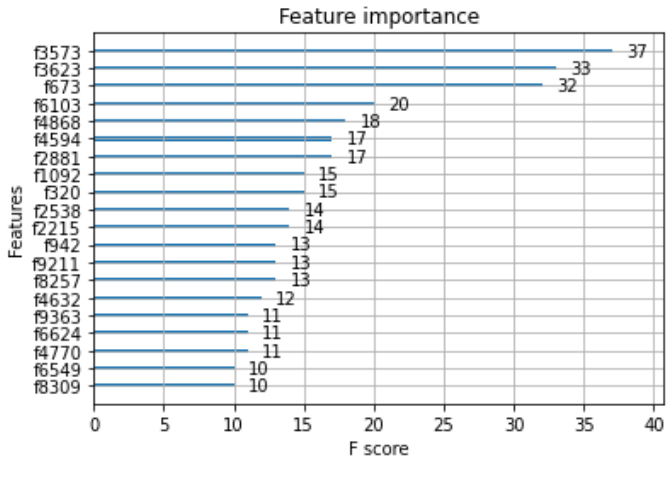
- DNN
  - model imfomation

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dropout_2 (Dropout)          (None, 29640)             0
_________________________________________________________________
dense_4 (Dense)              (None, 500)               14820500
_________________________________________________________________
activation_4 (Activation)    (None, 500)               0
_________________________________________________________________
dropout_3 (Dropout)          (None, 500)               0
_________________________________________________________________
activation_5 (Activation)    (None, 500)               0
_________________________________________________________________
dense_5 (Dense)              (None, 2)                 1002
_________________________________________________________________
activation_6 (Activation)    (None, 2)                 0
=================================================================
Total params: 14,821,502
Trainable params: 14,821,502
Non-trainable params: 0
_________________________________________________________________
```

  - add Dropout layer
  - no too deep (2 hidden layers)
- XGBoost
  - We compare the importance of the features with and without POS and both not setting the limitations on max features, and then we sort the most important twenty features.



- **Without POS：**



- **With POS：**
  - In the comparison of the two pictures, we found that in the without-POS model, the features with greater importance have higher F-score than those in the POS model. That means the words we've deleted in the POS tagging process is useful for us to achieve our target, which is to predict whether it is positive or not.

Discuss and answer the question

requirement :

- pdf file (1-2 pages, show your comparison)
- What can you conclude after performing tests 1, 2, 3? Please suggest the best model from the your team perspective.

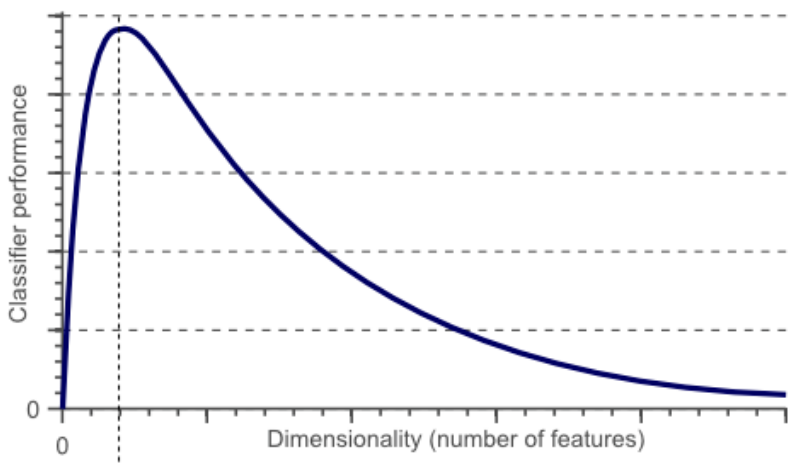
comparison :

- **Test 1: Lemmatization & POS v.s No Lemmatization & POS** We found that in the classification tree, data **without POS tagging** and lemmatization gives better result, and its accuracy is higher than that with POS tagging and lemmatization.
- **Test 2: Max features(500000,5000,3000,1000)** We choose the data without POS tagging and lemmatization, and then try four different number of max feature(500000,5000,3000,1000) to train the data. Finally, we find that **the largest features(500000)** gives best output.
- **Test 3: random forest ; classification tree ; neural network (default setting) ; neural network (user setting)** After testing different classification algorithm, we found that neural network (default setting) gives the best output, the accuracy is 0.842, and the F-score is 0.8381147540983608, and this process costs about 360 seconds to finish.

- **Conclusion**

  By concluding the test 1,2,3 in our model, we found that using neual network to classify by setting max feature=500000 without POS tagging will give us the best accuracy and F-score. We come up with some possible and reasonable explanations of this outcome. First of all, the reason why **without POS tagging gives better outcome** might be that it keeps more words, which **gives our model more features to select and analyze**, and we get higher accuracy. Besides, the amount of data is big enough; therefore, we can use max feature=500000 to analyze. Then by the testing, neural network with defualt setting has the highest accuracy, we consider that it's because its training way is more complicated and more accrate. Whereas, it also take the longest time to run, which is about 360 seconds, and if we **change its layers and activation algorithm together**, it will spend **over 3500 seconds** to finish; thus, if we are in hurry, the best choice would be the random forest. Moreover, we use two more classification algorithms(XGboost and DNN using keras) to analyze the data, which improve the accuracy up to 0.865.

  The reason why we set a limit of max features is to prevent **curse of dimensionality**, which refers to the performance of classifier is the highest with specific dimensions. When we have fewer amount of feature limits, the data with POS tagging won't have much change on its forecast ability toward the so-called general cases. The data without POS tagging originally have better predicting ability toward different special cases for it has more features as references, while when we have limitations on the amount of features, the left features for those special case have weaker abilities in predicting the general cases; thus the overall accuracy and F-score decrease. So that the smaller number of features can better analyze the data than the larger number of the features, which indicates that the curse of dimensionality does not happened with sufficient data in this case.



In conclusion, whether the text with POS tagging and how to set the feature amount to get the best accuracy depends on the amount and attributes of original data. In the case of this homework, we have large amount of data; thus it is better to use more features without POS tagging to train our data.